INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Contributions des techniques du traitement automatique des langues à la recherche d'information

Fabienne Moreau and Pascale Sébillot

N°5484

Février 2005

Systèmes symboliques

ISSN 0249-6399 ISRN INRIA/RR--5484--FR

Contributions des techniques du traitement automatique des langues à la recherche d'information

Fabienne Moreau * and Pascale Sébillot †

Systèmes symboliques
Projet TexMex

Rapport de recherche n ° 5484 — Février 2005 — 34 pages

Résumé : Les techniques issues du traitement automatique des langues (TAL) permettent de mettre au jour des informations morphologiques, syntaxiques et sémantiques sur les unités lexicales composant des textes. Ces divers types de connaissance ont été partiellement exploités par de nombreux travaux s'intéressant à l'interrogation de bases documentaires. Ce document, à travers un état de l'art de ces recherches, tente d'évaluer l'impact des différentes sortes d'informations linguistiques acquérables par des techniques de TAL sur les systèmes de recherche d'information et leurs performances, et de dresser un bilan de leurs contributions.

Mots-clé : Traitement automatique des langues, recherche d'information, indexation, extension de requêtes, informations morphologiques, syntaxiques et sémantiques.

(Abstract: pto)

* fabienne.moreau@irisa.fr
† pascale.sebillot@irisa.fr

Unité de recherche INRIA Rennes
IRISA, Campus universitaire de Beaulieu, 35042 RENNES Cedex (France)
Téléphone : 02 99 84 71 00 - International : +33 2 99 84 71 00
Télécopie : 02 99 84 71 71 - International : +33 2 99 84 71 71

Contributions of natural language processing techniques to information retrieval

Abstract: Natural language processing (NLP) techniques provide means to extract morphological, syntactic, and semantic information about lexical units in texts. This knowledge has been partly exploited by numerous works concerned by textual database questioning. This document presents a state-of-the-art of this field that mixes NLP and information retrieval (IR). It aims at evaluating the impact of linguistic information gleaned from NLP on the performances of IR systems that integrate them.

Key-words: Natural language processing, information retrieval, indexing, query expansion, morphological, syntactic, and semantic information.

Table des matières

1 Introduction

L'objectif d'un système (S) de recherche d'information (RI) est de retrouver, parmi une masse volumineuse de documents, ceux qui répondent précisément au besoin d'un utilisateur, besoin formulé par le biais d'une requête en langage naturel. La principale difficulté pour ces SRI est d'établir une correspondance entre l'information recherchée et l'ensemble des documents d'une collection. Pour y parvenir, ils se fondent généralement sur un appariement entre les mots contenus dans la requête et ceux, potentiellement pondérés, qui représentent le contenu de chaque document. La pertinence d'un document est alors évaluée en fonction des termes communs qu'il possède avec la requête.

Compte tenu de ce mécanisme de mise en correspondance basé sur une simple comparaison de chaînes de caractères, les SRI se trouvent rapidement confrontés à deux problèmes. Le premier concerne les formulations différentes d'un même concept : un document pertinent peut contenir des termes « sémantiquement proches » de ceux de la requête mais toutefois différents (synonymes, hyperonymes, termes ayant une forme morphologique différente...). Ce phénomène provoque une baisse du rappel de ces systèmes qui ne peuvent proposer à l'utilisateur certains documents pourtant intéressants. À ce problème vient s'ajouter celui de la polysémie des mots. L'ambiguïté qui en découle est à l'origine d'une baisse de précision des systèmes puisqu'elle entraîne potentiellement la récupération de documents non pertinents.

Pour faire face à ces difficultés liées à la complexité du langage naturel, une solution souvent évoquée est d'intégrer, au sein des SRI, une analyse linguistique qui présente l'avantage de ne plus considérer les mots comme de simples chaînes de caractères mais comme des entités linguistiques à part entière. Les traitements linguistiques en RI, effectués par le biais de techniques du traitement automatique des langues (TAL), extraient automatiquement des informations linguistiques des documents et des requêtes. Ces connaissances ont pour ambition de permettre aux SRI une meilleure compréhension des contenus et, par conséquent, d'avoir un impact sur leurs performances. Les traitements linguistiques peuvent intervenir de différentes façons dans un SRI. Ils contribuent, d'une part, en exploitant les connaissances linguistiques extraites des textes, à améliorer le processus d'indexation des documents, et à créer une représentation plus riche de leur contenu ; cette représentation vise à obtenir un appariement plus pertinent entre l'information recherchée par l'utilisateur et les documents de la collection. Ils ont, d'autre part, pour objectif d'améliorer le processus de recherche des systèmes en enrichissant les requêtes par des informations complémentaires, leur offrant ainsi la possibilité de retrouver davantage de documents intéressants.

Notre objectif, dans ce document, est de présenter une synthèse des contributions possibles des techniques issues du TAL pour une application de RI, à travers un tour d'horizon des diverses tentatives qui ont déjà été réalisées dans ce domaine. En TAL, on distingue généralement trois principaux niveaux d'analyse linguistique : les niveaux morphologique, syntaxique et sémantique. Pour parvenir à dresser un bilan assez exhaustif des différents apports possibles du TAL à la RI, nous choisissons de reprendre ce découpage et de nous attarder, pour chacun de ces niveaux, sur la façon dont les informations linguistiques extraites des documents et des requêtes ont été jusqu'à présent prises en compte par les systèmes.

Nous nous intéressons dans un premier temps au niveau morphologique de la langue. L'exploitation en RI d'informations morphologiques a pour objectif de permettre aux systèmes de reconnaître, au sein des documents et requêtes, les différentes formes d'un même mot et de pouvoir les apparier, limitant ainsi la baisse de rappel due à cette variation morphologique. Les techniques issues du TAL offrent différentes possibilités pour procéder à l'analyse morphologique des documents et requêtes, *i.e.* pour opérer une normalisation des formes des mots ou acquérir des informations morphologiques à partir des textes. Nous faisons un tour d'horizon des diverses façons d'exploiter ces informations et ces techniques en RI, et analysons comment il est possible de les intégrer au cœur même des systèmes. Nous cherchons à évaluer concrètement l'impact de l'analyse morphologique sur les performances des systèmes tant en termes de rappel que de précision.

Le second niveau de langue exploré est celui de la syntaxe. L'intégration de ce niveau dans l'analyse des documents et requêtes peut en effet s'avérer pertinent pour plusieurs raisons. D'une part, ceci conduit à extraire des documents de nouvelles entités linguistiques – plus précisément des termes complexes ou structurés – qui peuvent avoir une influence directe sur la résolution d'ambiguïtés, ces entités étant généralement moins polysémiques que les termes simples habituellement pris en compte par les systèmes. Ceci permet d'autre part, de pallier une des faiblesses des SRI, liée à leur représentation du contenu textuel dite en « sac de mots », en repérant et exploitant les différentes relations et dépendances qu'entretiennent les termes. La plupart des systèmes s'appuient, en effet, sur un modèle[1] où les documents sont représentés sous la forme d'un ensemble de termes non structurés, indépendants les uns des autres et non classés. L'objectif d'une vision plus structurée du document

1. *E.g.* les modèles vectoriel, booléen...

est de manipuler des informations plus précises et de faire référence aux concepts contenus dans les documents [CH01]. Une fois ces termes et ces structures extraits, la principale difficulté réside toutefois dans leur prise en compte effective au sein des SRI. Cela impose en effet une intégration dans des modèles de représentation où la notion de dépendance entre termes est en contradiction avec leur indépendance généralement sous-jacente ; cela implique également l'adaptation de mécanismes de fonctionnement, tels que, par exemple, la transformation des mesures de pondération pour prendre en compte les termes complexes. Nous faisons donc état, à travers la description de quelques travaux, des solutions qui ont été proposées pour répondre à ces problèmes et des différentes stratégies testées.

Cet article se termine par un focus sur la prise en compte du niveau sémantique de la langue dans les SRI. L'intégration d'informations sémantiques (acquises automatiquement à partir de textes ou issues de ressources existantes) peut se faire au niveau de l'indexation. Certains travaux développent ainsi une indexation dite « sémantique », qui cherche à associer aux termes d'un index un ensemble de sens non ambigus. Ce type d'indexation doit donc nécessairement être couplé à un traitement de désambiguïsation performant, et nous analysons les différentes propositions faites en ce sens. L'intégration de connaissances sémantiques se situe aussi au niveau de l'extension des requêtes de l'utilisateur. Les requêtes sont alors enrichies grâce à l'ajout de mots liés sémantiquement à ceux qu'elles contiennent, ce qui permet d'élargir le champ de recherche. Ce traitement nécessite également, comme nous le montrons, un processus de désambiguïsation des termes, afin de contrôler que le sens des mots utilisés pour l'extension soit identique à (ou proche de) celui des termes de la requête.

Avertissement : Le nombre de travaux ayant cherché à intégrer de façon souvent partielle des ressources linguistiques en RI est considérable. Cet état de l'art ne peut donc en faire une liste totalement exhaustive, et nous sommes conscientes que le lecteur avisé du domaine pourra certainement noter l'absence de telle ou telle référence, aussi pertinente que certaines que nous avons citées. De même, il ne vise pas, par les travaux qui y sont mentionnés, à faire une présentation explicite de chacune des expériences menées, la lecture des références données au fil du texte permettant de combler cette description partielle. Son objectif est de tenter de mettre en évidence les idées-clés de ce couplage TAL et RI, afin de comprendre les limitations exactes des modèles de SRI actuels pour la prise en compte de ressources linguistiques, et préciser les efforts qu'il reste à réaliser sur ces modèles pour accroître leur perméabilité à ces ressources.

2 Prise en compte de connaissances morphologiques en recherche d'information

Après un bref rappel de quelques notions fondamentales en morphologie, nous nous intéressons, dans cette section, au traitement de la variation morphologique en RI. Nous faisons état des différents outils d'analyse qu'il est possible d'appliquer sur les documents et les requêtes pour reconnaître et prendre en compte ces variantes, en distinguant ceux qui sont seulement une approximation de traitements linguistiques (racineurs [2]) et les analyseurs plus sophistiqués qui procèdent à une analyse linguistique plus fine des documents et requêtes (analyseurs flexionnels et dérivationnels [3]). Nous terminons par la description d'expériences qui comparent l'efficacité de ces deux techniques, afin de tenter de déterminer celle qui a le plus d'impact sur les performances des SRI.

2.1 Quelques notions utiles de morphologie

La morphologie [Pol03] concerne la structure des mots, c'est-à-dire les combinaisons de morphèmes (*i.e.* plus petites unités de sens) qui les forment. Ces morphèmes sont soit lexicaux, soit grammaticaux (les affixes), ces derniers se combinant aux précédents. La morphologie flexionnelle étudie les flexions des mots (marques de genre, nombre, personne, temps...) ; elle permet de relier les unités *chien* et *chiens* par exemple. La morphologie dérivationnelle s'intéresse à la formation de nouvelles unités lexicales à partir de morphèmes lexicaux et d'affixes dits dérivationnels ; elle explique la formation de *chanteur* à partir de *chant* et du suffixe *eur*. La morphologie compositionnelle, qui ne sera pas abordée ici, porte également sur la formation de nouvelles unités, mais par juxtaposition de plusieurs unités [Gro96].

Un mot peut donc avoir plusieurs formes qui possèdent un sens proche (*e.g. transformer*, *transforme*, *transformateur*, *transformation*...), et il n'est parfois pas pertinent – c'est le cas, bien souvent, en RI – de les distinguer. Il convient par conséquent de déterminer les diverses formes effectives d'un même mot, ou sa famille morphologique [4], par exemple en l'extrayant automatiquement de documents. Ceci permet alors de procéder à leur normalisation, c'est-à-dire à un recodage des diverses variantes du mot par une forme unique. Ce sont les possibilités offertes par l'analyse morphologique.

Si l'on s'oriente vers la morphologie flexionnelle, la forme unique sera le lemme (*i.e.* une forme canonique débarrassée de ses flexions). Le traitement associé à ce type de morphologie est la lemmatisation qui consiste à reconnaître, pour chaque mot, sa forme de base en supprimant ses traits de morphologie flexionnelle. Le recours à la morphologie dérivationnelle aura pour principal résultat la normalisation des formes autour d'une racine [5] ou d'un radical (*i.e.* une des formes prises par la racine [6]). Il existe, là aussi, différentes approches pour prendre en compte ce type de morphologie, comme le recours à des ressources morphologiques spécifiques ou l'utilisation d'analyseurs dérivationnels [7] [DFS02].

Enfin, pour traiter la variation morphologique des unités lexicales, on trouve également des approches moins sophistiquées (considérées comme des approximations des traitements linguistiques) qui permettent de rassembler les différentes variantes d'un mot autour d'un *stem* (ou pseudo-racine) qui se rapproche de la notion de racine (sans nécessairement avoir une origine étymologique comme dans le cas de la racine linguistique). Le traitement permettant d'obtenir ce type d'information est la procédure de *stemming* (ou racinisation), qui présente généralement l'avantage d'être moins complexe à mettre en œuvre que les deux types de traitements précédents. Cette procédure, qui cherche à regrouper les mots appartenant à une même famille morphologique, peut se définir comme une technique de désuffixation et recodage supprimant les affixes (essentiellement les suffixes) pour isoler des pseudo-racines. Elle prend en compte à la fois les cas relevant des morphologies flexionnelle et dérivationnelle.

2. *Cf.* section 2.2.1.

3. *Cf.* section 2.2.2.

4. Une famille morphologique peut être définie comme regroupant un ensemble de mots possédant une racine morphologique commune, obtenue à partir d'une analyse flexionnelle et dérivationnelle.

5. « On appelle racine l'élément de base, irréductible, commun à tous les représentants d'une même famille de mots à l'intérieur d'une langue ou d'une famille de langue. La racine est obtenue par élimination de tous les affixes et désinences. Elle est porteuse des sèmes essentiels, communs à tous les termes constitués avec cette racine » [Lar98].

6. La différence entre racine et radical est souvent ambiguë. La racine est la forme abstraite servant de base de représentation à tous les radicaux qui en sont les manifestations. Ainsi, la racine /chant/ possède deux radicaux : *chant-* et *cant-*. Parfois, une racine ne possède qu'un radical. Dans ce cas, les termes racine et radical peuvent se confondre.

7. *I.e.* des analyseurs morphologiques s'appliquant aux formes dérivées.

2.2 Traitements de la variation morphologique en RI

Les différentes expériences réalisées pour évaluer l'intérêt de recourir à un niveau d'analyse morphologique des documents et requêtes en RI n'utilisent pas nécessairement les mêmes types de traitements morphologiques. Afin d'en dresser un bilan efficace, nous décomposons la description de ces expérimentations selon le type d'outils qu'elles manipulent. Nous nous attachons dans un premier temps à décrire l'apport en RI d'une procédure de *stemming*. Nous nous intéressons ensuite à l'impact des analyseurs morphologiques flexionnels couplés à des analyseurs dérivationnels.

2.2.1 Utilisation d'un racineur

Il existe deux façons principales d'utiliser une procédure de racinisation (ou *stemming*) dans un SRI. La première consiste à l'appliquer lors de la phase d'indexation (c'est généralement le cas dans toutes les expériences décrites plus bas). Les mots des documents et requêtes sont ramenés à leurs *stems*, et l'appariement entre les termes de la requête et ceux des documents s'effectue donc sur cette base. La racinisation peut aussi intervenir uniquement pour la tâche d'extension des requêtes. Ces dernières sont alors enrichies à l'aide de termes morphologiquement liés à ceux qu'elles contiennent, généralement par le biais de familles morphologiques.

Nous nous intéressons ici à cette procédure de racinisation et aux informations morphologiques (d'ordre flexionnel et dérivationnel) qu'elle permet d'acquérir lors de l'analyse des documents et requêtes. Nous présentons pour débuter quelques résultats d'expérimentations réalisées pour mesurer l'impact de *stemmers* sur les performances de SRI. Dans les deux sous-sections suivantes, nous dressons un premier bilan, en nous appuyant là aussi sur des expériences concrètes, ayant pour but d'une part, de distinguer les cas où le *stemming* est efficace des cas où il semble néfaste et, d'autre part, de déterminer les facteurs qui ont une influence directe sur l'efficacité de cette procédure.

Quelques résultats

Comme nous venons de le mentionner, la première série d'expériences que nous allons décrire a pour objectif d'évaluer l'apport en RI d'une procédure de *stemming* des documents et des requêtes. Les *stemmers* traditionnellement utilisés au sein de ces travaux sont ceux de Porter [Por80] et Lovins [Lov68].

L'impact des algorithmes de *stemming* sur les performances de SRI varie selon les expérimentations. Les expériences de Lennon *et al.* [LPTW81] et Harman [Har91][8] mesurent l'influence de la racinisation pour la langue anglaise. Elles aboutissent à des conclusions globalement décevantes puisqu'aucune amélioration de résultats n'est constatée par rapport à un SRI « traditionnel » (*i.e.* qui ne prend pas en compte les variantes morphologiques). Les mêmes observations sont également obtenues par Fuller et Zobel [FZ98], qui comparent l'apport respectif de quatre types de *stemmers* (les algorithmes de Porter, de Lovins, le *S-Stemmer* et un *stemmer* à base de dictionnaires). Leur conclusion est que les améliorations apportées par ce traitement restent insuffisantes, bien qu'elles existent pour certaines requêtes.

Dans une étude plus détaillée, Hull [Hul96] tente de justifier les faibles résultats obtenus lors des expériences de Harman et Lennon *et al.* Il montre ainsi que les mesures d'évaluation traditionnelles de la RI (*i.e.* le rappel et la précision) ne sont pas forcément appropriées pour évaluer précisément l'influence du *stemming*. Il propose de nouvelles mesures[9], et les expériences réalisées[10], en revisitant l'algorithme de Porter (dictionnaires ajoutés pour valider les résultats après élimination des suffixes), montrent que le *stemming* est efficace pour l'Anglais, sauf pour les requêtes longues. La hausse des performances de recherche est sensible pour les requêtes courtes, le *stemming* permettant entre 1 et 3% d'amélioration par rapport aux systèmes ne prenant pas en compte les variantes morphologiques. Cette amélioration n'est toutefois pas effective sur de très petites collections de documents.

D'autres expériences relatives à l'utilisation du *stemming* en RI apparaissent également encourageantes. Pour Krovetz [Kro93], la racinisation conduit à une hausse des résultats située entre 1,3% et 45,3% selon les collections et les techniques de *stemming* utilisées. Il montre de façon plus précise que l'accroissement de la précision moyenne est d'environ 40% pour les textes courts (environ 45 mots) associés à des requêtes courtes (comportant 7 mots en moyenne) contre seulement 2% pour les textes longs (environ 500 mots) couplés à des requêtes courtes (avec 9 mots en moyenne). Loupy *et al.* [LBEM98] proposent aussi une série d'expériences concernant

8. Harman compare à la fois les *stemmers* de Porter, de Lovins et son *stemmer* (fondé sur un fonctionnement basique qui consiste à supprimer seulement quelques terminaisons (principalement les marques du pluriel)) lors de recherches dans de petites collections de documents (tests réalisés sur les collections Medlars, Cranfield et Cacm).

9. Ces mesures (basées notamment sur le test de Friedman) considèrent, entre autres, la longueur des requêtes et le nombre de documents pertinents retournés pour chacune d'entre elles.

10. Sur une large collection de documents (180000).

plus particulièrement l'enrichissement des requêtes. Les résultats obtenus illustrent l'intérêt du *stemming* pour cette tâche : une amélioration de 3,6% de la précision est constatée pour l'Anglais.

Il ressort de ces différentes expériences que l'influence du *stemming* en RI est tributaire d'un certain nombre de facteurs. Ainsi, alors que pour certaines requêtes, la racinisation s'avère très bénéfique [Hul96], elle peut entraîner, pour d'autres requêtes, une dégradation des performances. Il est donc nécessaire d'être capable de distinguer les cas où le *stemming* est intéressant des cas où il ne l'est pas, ce qui demeure toutefois une tâche difficile, comme nous allons le voir à présent.

Analyse de cas d'efficacité ou de non efficacité de la racinisation

Quelques travaux ont proposé des stratégies pour tenter d'analyser plus précisément les cas où la procédure de *stemming* s'avère efficace. Une première tentative d'explication proposée par Harman [Har91] a consisté à penser que le *stemming* aurait plus d'impact s'il était appliqué uniquement pour étendre les termes de la requête les plus discriminants[11]. Cependant, aucune amélioration de résultats n'a été constatée lors des expérimentations réalisées. Dans une autre expérience, Harman a cherché à évaluer si une amélioration était possible en étendant une requête par le biais de variantes morphologiques choisies par l'utilisateur. Les tests montrent alors une hausse substantielle des résultats et impliquent que l'intervention d'un utilisateur pour opérer un filtre sur les termes à étendre au sein de sa requête peut s'avérer utile.

Parallèlement à ces observations, d'autres remarques [Chu95, Kro93] tentent d'expliquer l'irrégularité des résultats obtenus lors de l'application du *stemming* en RI. Elles reposent sur l'observation d'un lien entre l'efficacité des *stemmers* et la prise en compte d'un niveau sémantique pour relier les variantes morphologiques. En effet, traditionnellement les *stemmers* procèdent à la normalisation des formes morphologiques sans considérer le sens des mots à raciniser. Or, il apparaît que les requêtes qui retrouvent des documents avec le plus de précision sont celles dont les variantes morphologiques proposées possèdent un lien sémantique, ce qui dépend du type de *stemmer* utilisé et de la qualité de la racinisation effectuée. Une mauvaise racinisation[12] a pour conséquence de regrouper des variantes qui font référence à des concepts différents, ce qui entraîne une dégradation des performances d'un SRI l'utilisant. Il n'est toutefois pas toujours évident d'évaluer précisément le lien sémantique qui unit deux variantes, et ce lien peut aussi dépendre du domaine de connaissance des documents, ceci expliquant pourquoi les diverses études aboutissent à des résultats mitigés. L'intégration d'une procédure de *stemming* au sein d'une stratégie de recherche qui prend en compte le sens des mots est une solution possible pour l'amélioration des performances des SRI. Dans cette optique, nous pouvons citer l'approche mise en œuvre par Xu et Croft [XC98] qui consiste à appliquer une procédure de racinisation pour former des classes de mots qui sont utilisées pour l'extension de requêtes. De façon plus précise, la méthode proposée, qui fonctionne en corpus, sans ressources ni règles prédéfinies, consiste dans un premier temps à normaliser les mots à l'aide d'un *stemmer* du type Porter. Elle rassemble ensuite au sein d'une même classe d'équivalence morphologique[13] les mots racinisés qui, en plus de posséder la même pseudo-racine, cooccurrent de façon significative (valeur mesurée à l'aide d'une variante de l'information mutuelle). L'utilisation des cooccurrences présuppose l'idée d'une proximité sémantique entre termes. Les résultats obtenus lors de ces expériences montrent une amélioration significative des performances.

Il semble donc, à travers ces différentes observations, que la procédure de *stemming* doive être couplée à des traitements complémentaires (comme la prise en compte d'une dimension sémantique) si l'on souhaite qu'elle ait un impact plus important en RI.

Analyse des facteurs contribuant à un meilleur impact des *stemmers*

De nombreux facteurs entrent en fait en jeu dans l'évaluation précise que l'on peut faire de l'impact du *stemming*. Comme nous l'avons vu, les résultats dépendent de la requête, longue ou courte, de l'utilisateur. Il est également nécessaire de prendre en compte le type de *stemmer* utilisé : un taux d'erreurs important introduit nécessairement un bruit élevé et donc une baisse de la précision du système. Pour réduire ces erreurs, Krovetz [Kro93] a proposé une nouvelle approche du *stemming* basée sur l'utilisation de dictionnaires électroniques comme source additionnelle de connaissances. L'implémentation de son outil (KSTEM) consiste à associer à un *stemmer* traditionnel du type Porter un dictionnaire électronique de l'Anglais. Ce dictionnaire exerce un contrôle et stoppe le processus de suppression de suffixes lorsque le mot obtenu y est trouvé. Bien que son

11. Termes de fréquence moyenne dans les documents.

12. Une mauvaise racinisation est généralement provoquée par une procédure de désuffixation excessive ou trop souple qui entraîne des erreurs dites de sur-racinisation ou de sous-racinisation. La sur-racinisation implique que la pseudo-racine est trop large (*e.g.* la pseudo-racine *nat* qui regroupe à la fois *nature* et *nation*). Inversement, pour la sous-racinisation, la pseudo-racine n'est pas assez large (*e.g.* la pseudo-racine *adaptat* qui empêche le regroupement des formes *adapter* et *adaptation*).

13. *E.g. stocks*, *stock*, *stocked*, *stocking*...

stemmer règle en grande partie les problèmes de l'algorithme de Porter, les résultats en termes de rappel et de précision des SRI l'utilisant ne sont néanmoins pas améliorés pour autant. D'autres expériences relativisent également l'apport en RI d'un *stemmer* basé sur des dictionnaires. L'expérimentation menée par Fuller et Zobel [FZ98] évalue les performances de différents *stemmers* (*stemmers* traditionnels et à base de dictionnaires) en mesurant leur capacité à retrouver les différentes variantes morphologiques d'une même forme. Il apparaît que le *stemmer* basé sur l'utilisation d'un dictionnaire fournit généralement la meilleure précision mais passe à côté de beaucoup de termes (*i.e.* des termes qui auraient dû être rapprochés car étant morphologiquement liés) qui sont absents de la ressource. De façon plus précise, les résultats observés montrent que le *stemmer* à base de dictionnaire retrouve correctement 48% des variantes morphologiques, contre 67% pour l'algorithme de Porter et 75% pour celui de Lovins. Il semble donc que le recours à de tels *stemmers* ne constitue pas une solution vraiment pertinente pour réduire le nombre d'erreurs. D'ailleurs, il est important de noter (comme le montrent Dal et Namer [DN00]) qu'il n'a jamais véritablement été démontré à quel point les erreurs générées par les systèmes de *stemming* pouvaient être néfastes aux performances des SRI. Les préjudices commis dépendent notamment du nombre (*i.e.* de la fréquence) de ces erreurs et également de leur contexte (par exemple, dans le cas de requêtes longues, la co-présence de nombreux mots dans la requête permet de minimiser les problèmes).

Une autre information importante à considérer lorsque l'on applique une procédure de *stemming* en RI est le type de forme qui est fournie en sortie du *stemmer*. Pour l'algorithme de Porter par exemple, le traitement produit un *stem* (*i.e.* pseudo-racine) qui peut être un mot ou seulement une partie de mot (*e.g. déménag-*) qui n'a pas toujours d'existence à part entière. Or, ce type de forme sera difficilement manipulable si l'on souhaite effectuer d'autres traitements plus complexes, nécessitant, par exemple, de confronter le *stem* obtenu à des données présentes dans des ressources (*e.g.* des dictionnaires). L'algorithme de Lovins quant à lui recode les formes obtenues afin d'obtenir des mots.

Enfin, un autre paramètre important est la langue sur laquelle le *stemming* est appliqué. Comme le soulignent Arampatzis *et al.* [AVKV00], l'efficacité du *stemming* dépend de la complexité morphologique de la langue. Les expériences réalisées sur des langues morphologiquement plus riches que l'Anglais montrent qu'une amélioration des performances de recherche peut être apportée par la racinisation. Popovic et Willett [PW92] constatent par exemple que le *stemming* en Slovène améliore de façon significative la précision (une augmentation d'environ 40% pour de petites collections de résumés). Pour cette langue, le *stemmer* développé prend en compte plus de 5200 suffixes. Pour le Suédois, Carlberger *et al.* [CDHK01] montrent que la procédure de *stemming* est intéressante en RI, puisqu'une augmentation du rappel et de la précision comprise entre 15 et 18% est observée pour des textes d'une longueur moyenne de 181 mots. Il apparaît donc que plus une langue est morphologiquement riche, plus il devient pertinent de prendre en compte le niveau morphologique de la langue à travers une procédure de *stemming*, voire de traitements morphologiques plus évolués.

En conclusion, nous retenons que l'utilisation d'une racinisation en RI peut être efficace à condition de tenir compte d'un certain nombre de facteurs. Nous pensons plus particulièrement à la nécessité, lorsque l'on regroupe les variantes morphologiques d'un même terme, de s'assurer de l'existence d'une relation entre ces termes qui dépasse le niveau morphologique (*i.e.* essayer de déterminer un lien sémantique). Il est également important de vérifier la fiabilité de l'outil utilisé (*i.e.* contrôler qu'il ne génère pas un taux d'erreurs trop important). Ce contrôle est d'autant plus important lorsque l'on utilise cette procédure sur des langues morphologiquement riches. Enfin, il est important de remarquer que la différence entre les résultats des expériences s'expliquent également par la difficulté à évaluer concrètement l'impact de ce traitement sur les performances des SRI, les mesures de rappel et de précision étant, en particulier peu adaptées et trop imprécises pour cette évaluation [Hul96, Pai94]. C'est pourquoi, avant de conclure définitivement sur l'apport ou non de la procédure de *stemming* en RI, il est certainement nécessaire de mener de nouvelles expériences qui utilisent des mesures d'évaluation plus appropriées.

Afin d'avoir une vision générale de l'intérêt des informations morphologiques au sein des SRI, nous nous intéressons à présent aux systèmes qui intègrent une analyse morphologique plus sophistiquée des documents et requêtes, analyse qui s'effectue par le biais de lemmatiseurs (qui se focalisent sur des informations de morphologie flexionnelle) et d'analyseurs dérivationnels (qui considèrent les informations de morphologie dérivationnelle). Notre choix de ne pas faire de distinction dans notre présentation entre ces deux types d'outils se justifie par le fait que, au sein des diverses expériences que nous allons décrire, ces traitements sont souvent couplés en raison de leur complémentarité.

2.2.2 Utilisation d'analyseurs morphologiques flexionnels et dérivationnels

Comme pour la procédure de racinisation, les analyseurs morphologiques (flexionnels et dérivationnels) interviennent à deux niveaux d'un SRI. Ils peuvent, en effet, être appliqués pour l'indexation des documents et des requêtes. C'est le cas le plus fréquent. L'appariement entre les termes de la requête et ceux des documents s'effectue donc sur la base du lemme (pour les analyseurs flexionnels) ou de la racine (pour les analyseurs dérivationnels). Ils peuvent aussi intervenir uniquement pour la tâche d'extension des requêtes. Il s'agit alors d'identifier les lemmes des termes des requêtes et d'enrichir ces dernières à l'aide des familles morphologiques constituées à partir d'analyses flexionnelles et dérivationnelles.

Nous nous intéressons ici à ces informations obtenues par le biais d'outils de lemmatisation et d'analyse dérivationnelle, appliqués lors de l'analyse morphologique des documents et requêtes. Après avoir donné certains résultats d'expérimentations réalisées pour évaluer leur impact, nous présentons quelques points-clés liés à l'application de ces traitements en RI.

Quelques résultats

Les expériences de Gaussier *et al.* [GGS97] pour le Français montrent l'apport de la morphologie flexionnelle en RI. Les résultats obtenus lors de l'intégration d'un module de lemmatisation[14] dans leur SRI présentent une amélioration de la précision moyenne de 16%. Ces auteurs proposent également de combiner le traitement de lemmatisation à un module dit de morphologie relationnelle[15]. Le fonctionnement de ce module, basé sur un lexique dérivationnel, est d'extraire des suffixes potentiels qui sont ensuite réutilisés pour tenter de relier les lemmes de même famille. Ce traitement additionnel offre une augmentation complémentaire de la précision moyenne de 2%.

Zweigenbaum *et al.* [ZGD01] montrent également l'apport faible mais réel de la lemmatisation et de la dérivation dans une tâche d'appariement entre requêtes et termes normalisés. La principale différence avec l'expérience précédente est que le SRI utilisé est spécialisé dans un domaine de connaissance restreint (*i.e.* le domaine médical). L'utilisation de connaissances flexionnelles[16] et dérivationnelles améliore en moyenne les réponses à une requête. La flexion agit dans 6,6% des cas avec une hausse modeste, et la dérivation agit dans 2% des cas avec une augmentation plus nette.

Les expériences de Vilares Ferro *et al.* [VBA02] pour l'Espagnol exploitent également successivement ces deux types d'analyseurs. La première étape du système consiste à étiqueter morpho-syntaxiquement les unités lexicales et à obtenir les lemmes des textes à indexer. Chacun des lemmes est ensuite remplacé par le représentant de la famille morphologique à laquelle il appartient. Une hausse significative du rappel peut alors être constatée par rapport à une procédure de *stemming* traditionnelle.

Points-clés

Au vu de ces expériences, il apparaît donc que l'analyse morphologique flexionnelle et dérivationnelle des documents et requêtes peut être considérée comme une tâche pertinente en RI. Grâce aux règles linguistiques fortes qu'elle met en jeu, elle limite les rapprochements de termes non liés. De plus, son application peut constituer une première étape de désambiguïsation des mots[17]. La prise en compte des variantes allomorphiques favorise le rappel. Néanmoins, comme pour la procédure de *stemming*, l'impact d'une analyse morphologique (flexionnelle ou dérivationnelle) semble étroitement lié à la langue prise en compte ; les expériences décrites précédemment ont cependant toutes été menées sur des langues morphologiquement riches (aussi bien le Français que l'Espagnol), et il convient d'étudier sa pertinence pour d'autres langues, telles que l'Anglais par exemple.

La combinaison des deux types de morphologie paraît être une solution efficace en vue de l'accroissement des performances des systèmes. L'intérêt du couplage est justifié, en effet, par la complémentarité des deux analyses : la lemmatisation permet, sans générer un nombre d'erreurs important, le regroupement de variantes morphologiques de même catégorie grammaticale. L'analyse dérivationnelle conduit ensuite, à partir des résultats de la lemmatisation, à rassembler (de façon plus précise qu'une procédure de *stemming* car motivée linguistiquement) les variantes morphologiques quelle que soit leur catégorie grammaticale. Il semble toutefois important, lors de leur intégration au sein d'un SRI, de bien cloisonner les traitements liés à la morphologie dérivationnelle des traitements liés à la morphologie flexionnelle afin d'optimiser au mieux leur efficacité, ce que confirment les expériences réalisées par Savoy [Sav93].

14. Utilisation des outils de lemmatisation de Xerox.

15. Nous considérons ici le terme relationnel comme équivalent à dérivationnel. Pour une distinction plus précise de ces notions, se référer à [GGS97].

16. L'analyseur flexionnel utilisé est le lemmatiseur FLEMM développé par Namer [Nam00].

17. L'analyse et la production de la forme de base d'un terme polysémique peut nécessiter la détermination automatique de sa catégorie morpho-syntaxique (*e.g. porte* = nom ou verbe) en se basant sur son contexte d'apparition.

La présentation de ces outils de lemmatisation et d'analyse dérivationnelle, telle que nous venons de la faire, ne nous permet cependant pas de mettre clairement en évidence leur valeur ajoutée par rapport à la procédure de *stemming* décrite précédemment. Il est donc nécessaire d'aller plus loin dans cette étude si l'on souhaite déterminer le type d'outils qui a le plus d'impact pour traiter le problème de la variation morphologique en RI.

2.3 *Stemming vs* analyse morphologique

Nous faisons ici état de différentes expériences qui comparent l'efficacité des *stemmers* à celle de lemmatiseurs et analyseurs dérivationnels. Les résultats de ces travaux nous permettent de dresser un bilan plus général qui montre l'intérêt d'évaluer l'influence des erreurs générées par ces outils, et met en valeur l'idée que le choix de l'outil à appliquer dépend fortement de la nature des traitements qui succèdent à l'analyse morphologique.

2.3.1 Expériences comparatives

Hull et Grefenstette [HG96] s'intéressent au calcul de la pertinence des résultats d'un SRI en réponse à une requête en comparant à des données témoins (*i.e.* requêtes et documents en Anglais non racinisés ni lemmatisés) celles traitées par quatre algorithmes : un algorithme de *stemming* – soit de Lovins, soit de Porter –, une procédure de lemmatisation et un module de dérivation. Leurs expérimentations indiquent que les performances des algorithmes dépendent des caractéristiques de la requête formulée [Nam00]. Il apparaît toutefois que pour des requêtes types (*i.e.* importantes relativement au test de Friedman), la lemmatisation est dans 40% des cas l'une des deux meilleures méthodes pour retrouver et classer les documents pertinents. Les auteurs analysent les cas d'échec comme étant liés au lemmatiseur utilisé dans l'expérience qui, d'une part, n'a pas la capacité de reconnaître les mots inconnus et, d'autre part, repose sur des règles qui, bien qu'étant pertinentes en linguistique[18], sont dommageables en RI[19]. Ils prédisent que pour une langue comme le Français, les résultats seront sans doute encore meilleurs, à condition d'utiliser un lemmatiseur non tributaire d'un dictionnaire.

Moulinier *et al.* [MML00] comparent à la fois les impacts respectifs d'un traitement de lemmatisation, d'une procédure de *stemming* (de type Porter) et d'un système sans analyse morphologique sur les performances d'un SRI, et mettent en valeur les différences importantes observées pour le Français et l'Anglais selon l'analyse mise en œuvre. Les résultats obtenus pour le Français peuvent être rapprochés de ceux de Hull et Grefenstette [HG96] pour l'Anglais, excepté que la lemmatisation fournit des résultats légèrement meilleurs. Deux observations majeures issues de ces travaux sont qu'un *stemmer* de type Porter est trop agressif et produit un nombre important d'erreurs pour le Français, et que le lemmatiseur utilisé dans l'expérience, basé sur un lexique, ne traite pas tous les mots des textes, témoignant par-là même des limites de l'utilisation de telles ressources dans un cadre de RI.

Dans l'ensemble, ces différentes études tendent à montrer l'intérêt de prendre en compte les variantes morphologiques pour améliorer le rappel et la précision des SRI. Mais il reste toutefois encore difficile de choisir clairement le type d'outils (*i.e. stemmer vs* analyseurs flexionnels et dérivationnels) réellement le plus adapté. Comme nous déjà l'avons vu, la procédure de *stemming* peut être bénéfique à certaines conditions (*i.e.* selon la langue prise en compte, le type de *stemmer* utilisé et les erreurs qu'il génère) ; l'approche la plus efficace consiste à coupler un racineur avec une procédure de contrôle (*e.g.* introduire une dimension sémantique entre les variantes morphologiques) qui permet de vérifier que les variantes regroupées ont effectivement un lien morphologique. L'analyse morphologique d'ordre flexionnel, qui paraît d'ailleurs assez fiable, contribue également à améliorer les performances des systèmes, et semble plus particulièrement adaptée aux langues morphologiquement riches. Son couplage avec un traitement de morphologie dérivationnelle, en procédant au rassemblement des formes quelles que soient leurs catégories grammaticales, offre l'avantage de retrouver plus de variantes.

2.3.2 Impact des erreurs

Si l'on cherche à évaluer l'influence exacte des erreurs générées par ces divers outils, on constate que celle-ci est fortement liée à la façon dont les variantes morphologiques vont être utilisées. Le système développé par Xu et Croft [XC98] par exemple ne nécessite pas une analyse morphologique fine puisque les résultats obtenus sont couplés à des traitements statistiques qui opèrent un filtrage des erreurs obtenues par la procédure de *stemming*. Ainsi, pour reprendre l'exemple de Dal et Namer [DN00], que les mots *marmaille* et *marmite* soient considérés comme appartenant à une même famille de radical n'est pas si préjudiciable, car les chances de voir apparaître

18. L'exemple donné est celui du nom *optics* qui n'est pas relié à *optic* puisque ce dernier est un adjectif.

19. L'exemple cité (toujours par les auteurs) est celui de la requête *fiber optics* qui ne permet pas de retrouver des documents (pourtant pertinents) traitant de *fiber optic*.

ces deux mots ensemble au sein de documents sont assez faibles. À l'inverse, pour les systèmes qui n'utilisent pas de filtrage statistique, les erreurs générées par les *stemmers* peuvent devenir rapidement néfastes, pénalisant aussi bien la précision que le rappel.

2.3.3 Prise en compte des traitements des autres niveaux linguistiques

Le choix de la méthode à adopter pour gérer le problème des variantes morphologiques dépend également des traitements qui succèderont à l'analyse morphologique. Si l'on souhaite intégrer des analyses linguistiques complémentaires au sein du SRI (*i.e.* prise en compte des niveaux syntaxique et sémantique de la langue), l'utilisation d'une analyse morphologique plus sophistiquée s'avère nécessaire dans la mesure où les traitements suivants exploiteront les informations fournies en sortie de cette première phase. Ainsi, une analyse syntaxique peut, par exemple, exploiter les informations flexionnelles fournies par le lemmatiseur ; les catégories morpho-syntaxiques générées par l'analyse morphologique sont également utiles pour un traitement de désambiguïsation. De plus, si l'on souhaite recourir à des ressources lexicales ou sémantiques (*e.g.* des dictionnaires), il est indispensable de disposer du lemme qui, contrairement au *stem*, est la forme standard des entrées lexicales de ces ressources.

Inversement, l'apport réel des informations morphologiques sur les performances des SRI pourrait certainement être accru par l'intégration d'analyses plus poussées (*e.g.* exploitation du niveau syntaxique ou sémantique). Un certain nombre d'expériences montrent que des erreurs des traitements morphologiques sont liées à la non prise en compte des termes complexes (extractibles par une analyse syntaxique) ou la présence de termes morphologiquement ambigus (désambiguïsables par le biais d'informations sémantiques ou syntaxico-sémantiques). Comme l'attestent Arampatzis *et al.* [AVKV00] ou Krovetz [Kro93], ce sont les variations linguistiques qui dégradent nettement les résultats. Traiter la variation morphologique constitue une première étape, qui doit être complétée par la prise en compte de celles d'autres niveaux de la langue pour se faire une opinion plus précise de ce que les informations linguistiques peuvent apporter à la RI.

3 Prise en compte de connaissances syntaxiques en recherche d'information

Cette deuxième section est dédiée à l'étude et l'évaluation de l'intégration d'informations d'ordre syntaxique au sein de SRI. Nous présentons tout d'abord les types de connaissances syntaxiques qui peuvent être concrètement exploités, en nous limitant à deux d'entre eux : les termes complexes et structurés. Nous décrivons successivement, dans les deux parties suivantes, diverses expériences prenant en compte chacune de ces sortes de données, et analysons les résultats obtenus. Nous terminons enfin par la mention d'adaptations qu'il est nécessaire de prévoir pour les SRI si l'on souhaite intégrer de telles informations.

3.1 Informations syntaxiques et RI

L'analyse syntaxique s'intéresse à l'étude de la structure des phrases et des syntagmes. Son application à un document permet [20] notamment de prendre en compte l'ordre des mots dans une phrase, d'identifier les fonctions grammaticales des termes, de repérer des termes complexes ou des collocations... Dans un cadre de RI, toutes ces informations syntaxiques ne sont pas forcément exploitables [21], et nous avons choisi de nous intéresser plus particulièrement à deux types de structures : les termes complexes et structurés. Nous débutons cette section par les définitions successives de ces deux notions, et expliquons ensuite comment elles peuvent être exploitées en RI.

3.1.1 Terme complexe

Un terme complexe est une unité lexicale constituée d'au moins deux termes pleins, auxquels peuvent s'adjoindre des déterminants et prépositions. Très fréquemment, les termes complexes étudiés contiennent deux noms [22].

Les termes complexes ont l'avantage d'être moins ambigus que les simples, et sont souvent plus aptes à désigner des concepts puisqu'ils réfèrent généralement un domaine de connaissance spécialisé. Ceci leur confère un intérêt particulier en RI pour représenter le contenu sémantique de documents et requêtes. Ils sont toutefois plus difficiles à repérer car sujets à des variations (*cf.* section 3.1.3).

Les termes complexes peuvent être extraits de textes à l'aide de techniques d'acquisition exploitant soit des indices numériques (aspect fréquentiel), soit des indices structurels (aspect symbolique), ou combinant ces deux approches de l'extraction [Cla03]. Alors que les techniques numériques offrent la possibilité de retrouver des paires de mots [23] qui apparaissent de façon fréquente dans un corpus, liés de manière statistiquement significative, les méthodes symboliques permettent d'extraire des combinaisons de termes dont la structure syntagmatique est connue et syntaxiquement correcte. Chacune de ces approches possède ses points forts et faibles. Brièvement, le principal atout des méthodes numériques est de couvrir de manière exhaustive toutes les combinaisons possibles de termes dans une fenêtre textuelle déterminée ; néanmoins, en l'absence de prise en compte du contexte linguistique des termes, certaines combinaisons, valables statistiquement, ne sont pas toujours motivées aussi bien syntaxiquement que sémantiquement. Ces limites peuvent être comblées par les méthodes symboliques qui, en contrepartie, nécessitent généralement le recours à des connaissances *a priori* sur les termes [24].

3.1.2 Terme structuré

Nous désignons par terme structuré un ensemble de mots qui entretiennent des relations de dépendance. Plus précisément, un terme structuré est une structure composée d'un terme nommé « tête » (élément central

20. Selon les applications, *e.g.* la traduction automatique, la correction orthographique, la recherche d'information..., les besoins sont différents.

21. Certaines informations exigent, pour leur acquisition, une analyse des documents particulièrement profonde qui n'est pas toujours réalisable en RI puisque le temps de traitement peut être un facteur important à prendre en compte, de même que les qualité et variabilité rédactionnelles des documents.

22. Exemples de termes complexes : accident de voiture, pince à épiler, ingénieur chimiste... En Français, les structures du type *Nom Prép (Det) Nom* sont fréquentes.

23. Nous restreignons ici notre définition des termes complexes à des paires de mots.

24. En effet, selon les approches utilisées, il est nécessaire de disposer de patrons (morpho-)syntaxiques de termes complexes, de lister des catégories ne pouvant apparaître dans un terme, ou de fournir des séries d'exemples dont on peut inférer des patrons d'extraction.

du groupe) qui régit un ensemble de termes « modifieurs » qui sont en relation de dépendance avec lui[25]. Avec cette définition, notons que l'on peut voir les termes complexes comme un sous-ensemble particulier des termes structurés.

Nous considérons ici principalement les syntagmes comme termes structurés, syntagmes dont la nature exacte est liée à la catégorie de leur tête[26]. Un document (ou une requête) peut alors être représenté à l'aide de ces combinaisons de termes fortement structurées et non plus comme un simple ensemble de mots isolés[27], ce qui contribue, grâce à l'identification des structures, à une meilleure compréhension de son contenu sémantique.

Pour pouvoir caractériser ces syntagmes et leur structure, il est nécessaire de procéder à une analyse syntaxique (plus ou moins fine) des documents et requêtes. Son objectif est d'identifier dans les documents et les requêtes les groupes syntagmatiques (groupes nominaux, verbaux...) qui déterminent la structure syntaxique (souvent partielle) des phrases et permettent d'établir des relations de dépendance entre les termes simples, exploitables en RI selon les méthodes que nous présentons ci-dessous.

3.1.3 Utilisation des termes complexes et structurés et de leurs variantes

Avant de se pencher sur l'utilisation effective de termes complexes et structurés en RI, il convient de noter un problème essentiel lié à ces structures complexes : la variabilité. Pour pouvoir tirer profit de ces structures, il faut savoir reconnaître leurs diverses variantes. On distingue [Dai02] traditionnellement pour les termes complexes les variantes typographiques (*e.g. système expert* et *système-expert*), morpho-syntaxiques (*e.g. ulcère de la cornée* et *ulcère cornéen*[28]), syntaxiques (*e.g. artère coronaire* dans *artères fémorales, rénales et coronaires* ou encore *sécurité des réseaux* et *sécurité des données et des réseaux*)[29]... Les termes structurés sont également sujets à ce type de variations comme l'illustre l'exemple du syntagme *information retrieval* cité en note 25. La normalisation de ces syntagmes en une structure tête+modifieur permet de regrouper les variantes en une seule et même forme.

Bien que non appliqués au sein d'un SRI, les travaux de Jacquemin *et al.* [JKT97] montrent l'intérêt de prendre en compte ces variations pour représenter des textes. Deux expériences sont réalisées afin de comparer une indexation sans et avec le recours aux variantes des termes complexes[30]. La seconde expérimentation détecte environ 30% de termes supplémentaires, offrant une indexation des documents à couverture plus large. Notons que pour être intéressante, l'identification de variantes doit toutefois être effectuée sans introduire de bruit, et sans affecter les temps de traitement dans le cas d'utilisations en ligne.

Les structures complexes peuvent, comme les informations morphologiques évoquées précédemment, être prises en compte soit lors de la phase d'indexation des documents et des requêtes, soit pour l'extension de ces dernières. Les diverses variantes possibles d'un terme complexe ou structuré[31] peuvent être normalisées (*i.e.* ramenées à une forme standard). Cette forme unique est alors utilisée pour l'indexation, l'appariement entre les documents et la requête se faisant par son biais. L'objectif est de favoriser les documents dans lesquels les termes pleins formant la structure complexe entretiennent la même relation de dépendance que dans la requête. Une seconde technique consiste à étendre la requête par les variantes des termes (complexes ou structurés) qu'elle contient. Nous séparons dans la suite de notre exposé l'utilisation des termes complexes et la prise en compte des termes structurés en RI.

3.2 Prise en compte de termes complexes au sein d'un SRI

Un certain nombre de travaux ont tenté d'intégrer des termes complexes au sein de SRI. Ils aboutissent cependant, comme nous allons le voir ici, à des résultats globalement mitigés, et il paraît difficile de tirer des conclusions claires quant à leur intérêt. On constate toutefois une certaine disparité de ces résultats selon la méthode d'extraction des termes complexes retenue : approche à base d'indices numériques ou structurels. Nous allons donc nous attarder successivement sur les expériences menées sur la prise en compte en RI de termes complexes extraits par l'une ou l'autre de ces techniques. Pour faciliter la lecture, nous choisissons d'appeler

25. Ainsi, dans un exemple emprunté à Strzalkowski *et al.* [SLWPC99], les syntagmes *information retrieval*, *retrieval of information*, *retrieve more information* et *information that is retrieved*... peuvent tous être ramenés à une même forme : la paire *retrieve+information* où *retrieve* est la tête et *information* est le modifieur.

26. La tête peut être un nom – on parle alors de syntagme nominal –, un verbe (syntagme verbal), un adjectif (syntagme adjectival), un adverbe (syntagme adverbial) ou une préposition (syntagme prépositionnel).

27. C'est la définition de la représentation en « sac de mots ».

28. Cet exemple est emprunté à Daille [Dai96].

29. Pour une étude approfondie sur les variantes, se référer aux travaux de C. Jacquemin [Jac94].

30. L'identification des variantes se fait par le biais du système FASTR [Jac94].

31. Contenues aussi bien dans les documents que dans les requêtes.

par la suite « terme complexe statistique » un terme obtenu par l'emploi d'une méthode d'acquisition se basant sur l'aspect fréquentiel des textes, *i.e.* numérique, et « terme complexe syntaxique » un terme complexe extrait par une technique reposant sur des indices structurels, *i.e.* symbolique. Nous terminons cette section en tentant de mettre en évidence les points importants issus de ces expérimentations.

3.2.1 Les termes complexes « statistiques »

Les travaux précurseurs réalisés dans la prise en compte de termes complexes « statistiques » en RI sont ceux de Salton *et al.* [SYY75]. L'intégration des termes extraits par repérage de cooccurrences lors de la phase d'indexation de documents montre une amélioration de la précision moyenne comprise entre 17 et 39%[32] selon les collections de documents utilisées[33]. Ces résultats sont réévalués par Fagan [Fag87], qui reprend les expériences de Salton *et al.* et élargit le nombre de documents et requêtes afin de procéder à une évaluation à plus grande échelle. La hausse de la précision moyenne est plus limitée, comprise entre 2,2 et 22,7%.

Malgré l'aspect plutôt positif des résultats de ces expérimentations, les améliorations observées sont modulées par deux facteurs. Le premier concerne les collections de documents traitées. Pour Fagan [Fag87], cette variation est caractéristique des limites de l'approche numérique d'acquisition de termes complexes. Il est, en effet, souvent difficile de faire la distinction entre terme et non terme (au sens terminologique). Puisque l'extraction numérique repose essentiellement sur la notion de fréquence mais sur aucun critère linguistique, elle génère toutes les paires de mots possibles, qui ne sont pas forcément syntaxiquement (et sémantiquement) correctes[34], ce qui peut induire un taux de bruit important et entraîner une baisse des performances. Le second facteur de variation est la langue de la collection. Ceci est illustré par exemple par les expériences de Gaussier *et al.* [GGHR00] qui consistent à évaluer, pour le Français, l'impact des termes complexes – plus précisément de paires adjacentes[35] – dans les index, en complément de termes simples. Sur leur corpus d'évaluation[36] et avec un jeu de requêtes très limité (11 questions), les améliorations constatées sont insuffisantes, et ne permettent pas de prouver l'intérêt des termes complexes en RI. Ces auteurs proposent alors d'ajouter à ces termes complexes un nouveau type de composés très fréquents en Français : les termes *Nom Prép Nom*. Leur prise en compte ne conduit toutefois qu'à une hausse de la précision moyenne limitée à 0,7%, et les scores de rappel et de précision restent proches de ceux obtenus avec les seules paires adjacentes.

3.2.2 Les termes complexes « syntaxiques »

Compte tenu des limites des termes complexes extraits statistiquement[37], un certain nombre de travaux a cherché à améliorer les résultats en prenant en compte des termes complexes extraits en se basant sur des indices structurels (essentiellement par exploitation de patrons prédéfinis). Nous nous intéressons donc à présent aux expériences qui évaluent l'efficacité de ce type de termes, et qui comparent leur impact à celui des termes complexes « statistiques ».

Suite à ses premières expérimentations ayant montré l'apport des termes complexes « statistiques » mais également leurs limites, Fagan [Fag87] mène une seconde série d'expériences exploitant les termes complexes « syntaxiques »[38]. Les résultats ne présentent aucune amélioration significative par rapport à une indexation par termes simples, et sont en dessous de celle à base de termes « statistiques ». Des conclusions identiques sont obtenues par Lewis [Lew92]. Gaussier *et al.* [GGHR00] poursuivent également les expérimentations citées précédemment, portant sur le Français, en s'appuyant sur un analyseur syntaxique[39] et des patrons syntaxiques prédéfinis. Là encore, les résultats demeurent inférieurs à ceux obtenus avec les paires adjacentes.

L'influence des termes complexes linguistiquement motivés semblent donc peu évidente. Cependant, pour ces trois expériences, les performances sont quelque peu biaisées, à la fois par le petit nombre de requêtes testées[40] et

32. Pour une pondération des termes de type *tf* (*term frequency*). Avec une pondération *idf* (*inverse document frequency*), l'amélioration est moins importante (entre 6 et 20%).

33. Dans cette étude, les collections utilisées sont Cacm, Inspec, Cran, Med et Cisi.

34. Pour illustrer cette idée, Fagan donne l'exemple de *parallel and sequential algorithms*. Alors que les méthodes statistiques peuvent générer les deux constructions *parallel sequential* et *sequential algorithms*, les méthodes symboliques génèrent des expressions plus pertinentes telles que : *parallel algorithms* et *sequential algorithms*.

35. *I.e.* toutes les paires de mots pleins contigus.

36. Utilisation des collections de documents du projet Amaryllis, fondées sur des corpus d'articles (environ 11000) issus du journal Le Monde.

37. Essentiellement liées au fait qu'ils ne sont pas toujours motivés linguistiquement.

38. L'identification de ces termes complexes se fonde sur l'utilisation d'un étiqueteur syntaxique (d'IBM) et de listes de patrons syntaxiques de termes complexes. Une normalisation morphologique des termes (par *stemming*) est également effectuée.

39. Analyseur syntaxique à base de transducteurs à états finis [HG96].

40. Un jeu d'une cinquantaine de requêtes paraît être un seuil minimal pour mener une expérimentation complète.

par les tailles encore trop petites[41] des collections de documents utilisées. Ces conditions sont particulièrement néfastes à l'évaluation de l'impact des termes complexes « syntaxiques ». La taille de la collection influe en effet sur la pondération accordée à ces termes peu fréquents, malgré leur véritable pouvoir de représentativité des contenus. Des conditions plus favorables sont donc envisageables.

D'autres tests montrent, en effet, l'intérêt d'exploiter les termes complexes « syntaxiques »[42]. Les expériences proposées par Hull *et al.* [HGS+97], qui ont pour objectif de déterminer les types de termes complexes les plus efficaces en RI, mettent en évidence la supériorité des termes complexes extraits par patrons syntaxiques (augmentation de la précision moyenne d'environ 15% par rapport à l'utilisation de groupes de mots extraits par une technique statistique). Une des explications de cette supériorité des méthodes symboliques est liée au fait qu'elles offrent, par le biais des patrons d'extraction, un contrôle de la pertinence des termes extraits. Dillon et Gray [DG83] montrent également l'apport des termes « syntaxiques ». Au sein du système FASIT, ils comparent une méthode d'indexation à base de termes simples (avec *stemming*) à une indexation par termes acquis par patrons syntaxiques. Les résultats obtenus conduisent à une hausse de la précision moyenne comprise entre 3 % et 7 % pour un rappel compris entre 40 et 60%.

3.2.3 Bilan : apport respectif des différents types de termes

Toutes ces expérimentations concluent de manière peu tranchée sur l'intérêt ou non de recourir à des termes complexes en RI, voire sur le mode d'extraction privilégié de ces termes. Quelques idées relevées au cours de deux sections précédentes peuvent toutefois être rappelées.

Les termes complexes extraits par des approches numériques peuvent contribuer à une amélioration des performances des systèmes dans certains cas. Leur efficacité dépend d'une part de la taille du corpus de données. En effet, plus le corpus de documents est volumineux, plus les méthodes numériques sont fiables puisqu'elles reposent essentiellement sur la notion de fréquence. Le problème de la pertinence est peut-être le facteur le plus important à considérer pour évaluer l'apport de ces termes en RI. L'hypothèse sur laquelle se fondent les méthodes numériques est que plus les termes sont statistiquement significatifs, plus ils auront de chance d'être sémantiquement pertinents et de contribuer par conséquent à une amélioration des performances. Les résultats mitigés obtenus s'expliquent par le fait que, dans certains cas, les termes repérés sont effectivement pertinents et sémantiquement significatifs, alors que dans d'autres cas de figure, ils ne fournissent aucun apport sémantique à la représentation du contenu des documents et des requêtes. De plus, des termes complexes peu fréquents mais pouvant être pertinents ne sont pas pris en compte par ces méthodes.

L'impact des termes extraits par une approche symbolique est aussi très variable. Ces termes offrent, en théorie, plus de possibilités d'accroissement de performances aux SRI. En effet, les méthodes symboliques permettent d'opérer, par l'intermédiaire des règles linguistiques mises en œuvre, un contrôle des éléments extraits, limitant le bruit. Il paraît donc nécessaire de procéder à de nouvelles expérimentations en utilisant des conditions d'évaluation qui soient plus favorables (*i.e.* accroître le nombre de documents et de requêtes...) pour mieux percevoir leur efficacité.

De façon plus générale, l'efficacité des termes complexes est donc dépendante de la qualité des éléments recueillis[43]. Elle est liée également à la façon de représenter ces termes dans les index. Ayant généralement une faible fréquence dans les documents, ils doivent nécessairement être bien pondérés[44] si l'on souhaite mettre en valeur leur importance (*i.e.* leur fort pouvoir de représentativité du contenu textuel). Différentes mesures ont été proposées et sont décrites en section 3.4.1.

Nous allons maintenant nous intéresser au second type d'informations syntaxiques exploitable en RI : les termes structurés.

3.3 Intégration de termes structurés au sein d'un SRI

Nous avons défini en section 3.1.2 les termes structurés comme un ensemble de termes liés par une relation de dépendance, généralement caractérisée par un élément tête qui régit un ou des termes modifieurs. Cette structure tête+modifieur peut être exploitée en RI de deux manières. Elle sert de représentation normalisée à toutes les variantes d'un même terme structuré (*cf.* section 3.1.3) ; la relation tête+modifieur est ensuite utilisée

41. Si on les compare aux collections constituées au cours de diverses campagnes d'évaluation Trec pour l'Anglais.

42. Nous faisons en particulier référence aux remarques de Sparck Jones [Spa99] qui indiquent la supériorité d'un point de vue théorique des méthodes motivées linguistiquement sur les approches non linguistiques pour la prise en compte des termes complexes.

43. Plus ils ont un fort pouvoir de représentativité du contenu textuel, plus ils ont un impact important sur les performances des systèmes.

44. Nous rappelons que dans les SRI traditionnels, le poids d'un terme, qui reflète son importance dans le document, est très souvent fonction de sa fréquence d'apparition.

pour permettre l'appariement ou étendre des requêtes. Elle peut également être exploitée lorsque l'on souhaite apparier des termes simples qui suivent une même relation. Ainsi, il est possible de chercher tous les termes *logiciel* et de ne retourner que les documents où ce terme est, par exemple, en position modifieur au sein d'un syntagme (*e.g. installation de logiciel, configuration du logiciel...*), laissant de côté ceux où il est en position tête (*e.g. logiciel de compatibilité et de gestion, logiciel client, logiciel de navigation...*).

Nous débutons cette section par la description d'expériences construisant et utilisant cette structuration des termes en paires tête+modifieur. Nous proposons ensuite quelques autres techniques pour prendre en compte la structure des termes en RI. Nous nous intéressons enfin à l'exploitation des termes structurés en post-traitement d'un SRI. Un bilan revient sur les résultats-clés de ces diverses expérimentations.

3.3.1 Structuration des termes en paires tête+modifieur

La principale méthode utilisée pour extraire les termes structurés consiste dans un premier temps à identifier au sein des documents et des requêtes les groupes de termes qui entretiennent des relations de dépendance, c'est-à-dire, plus simplement, de procéder à la reconnaissance des syntagmes (essentiellement nominaux[45]). Pour leur identification et la détection de leurs structures, les systèmes procèdent à une analyse syntaxique[46] et s'appuient sur un ensemble de patrons syntaxiques prédéfinis. Les syntagmes sont alors normalisés sous la forme tête+modifieur[47] qui explicite les dépendances unissant les deux termes. Ces paires tête+modifieur sont ensuite utilisées comme termes pour l'indexation, généralement en complément[48] des termes simples. Lors d'une interrogation, l'appariement consiste alors à comparer les termes structurés des documents à ceux de la requête.

Cette méthode a été expérimentée à plusieurs reprises. On peut citer les travaux de Strzalkowski *et al.* [SLWPC99] qui mettent en œuvre, pour l'identification des syntagmes (nominaux et verbaux), une analyse syntaxique robuste des textes[49] et effectuent une phase de normalisation en paires tête+modifieur (Nom+Adjectif, Verbe+Objet...). Les résultats obtenus montrent que l'emploi de ces paires a une influence légèrement favorable, plus perceptible pour les requêtes longues (augmentation de la précision moyenne de 4% pour les requêtes courtes contre 18% pour les longues). Dans son expérience, Haddad [Had02] effectue, en complément de la structuration des termes en tête+modifieur, un filtrage syntaxique des syntagmes extraits (qui consiste à privilégier les syntagmes nominaux les plus longs) afin d'affiner l'analyse et retenir uniquement les syntagmes pertinents (*i.e.* qui ont une structure correcte). Un accroissement net des performances du SRI testé est constaté, à la fois en termes de rappel et de précision (entre 5 et 30 % d'augmentation de la précision moyenne). Le système IRENA[50] développé par Arampatzis *et al.* [ATK96] propose, quant à lui, de compléter la représentation du contenu des documents, traditionnellement basée sur l'indexation par termes simples, par la prise en compte des expressions nominales et verbales structurées en relations tête+modifieur. Son originalité consiste à ajouter à l'analyse syntaxique un certain nombre de traitements statistiques, dont le but est d'identifier certaines collocations et de filtrer les syntagmes extraits syntaxiquement. Les résultats obtenus montrent que l'ajout de syntagmes nominaux et verbaux dans les index (des documents et requêtes) permet une amélioration de 5% de précision moyenne par rapport à l'indexation par termes simples.

Kraaij et Pohlmann [KP98] présentent une approche similaire à celle mise en œuvre par Strzalkowski *et al.*, mais pour le Néerlandais. L'ajout de paires tête+modifieur conduit à une augmentation des performances pouvant atteindre 25%. Des expériences ont également été réalisées pour l'Espagnol par Vilares Ferro *et al.* [VBA02], notamment pour prendre en compte les variantes des termes structurés. La méthode utilisée est, là aussi, proche de celle proposée par Strzalkowski *et al.*: les syntagmes nominaux sont extraits par une analyse syntaxique robuste et normalisés en paires tête+modifieur. Les résultats obtenus montrent une amélioration assez significative des performances, plus particulièrement en termes de précision. Pour le Français, les expériences de Faraj *et al.* [FGM+96] proposent également une méthode d'indexation basée sur l'exploitation des groupes nominaux[51] normalisés. Les expérimentations menées révèlent une amélioration systématique des performances (augmentation de 5,7% de la précision moyenne par rapport à une indexation par termes simples). Néanmoins, il est important de remarquer qu'elles s'appliquent sur un corpus spécialisé, composé d'un vocabulaire technique,

45. Certaines expériences prennent cependant en compte, comme nous le mentionnons par la suite, d'autres types de syntagmes.

46. L'analyse syntaxique est généralement basée sur un ensemble de règles qui prend en compte la structure des phrases et le contexte des termes. Elle permet également d'opérer l'étiquetage grammatical des termes présents dans les documents et les requêtes.

47. Ainsi, *information retrieval, retrieval of information, retrieve more information* et *information that is retrieved* sont normalisés et ramenés à une seule et même paire : *retrieve + information*.

48. Elles peuvent parfois être utilisées également en remplacement des termes simples.

49. L'outil syntaxique utilisé est l'analyseur TTP (Tagger Text Parser) basé sur les grammaires en chaînes de Sager [Sag81].

50. Information Retrieval system Engine based on Natural language Analysis.

51. Les groupes nominaux (plus précisément des synapsies) sont extraits à l'aide de l'outil TERMINO qui repose essentiellement sur une approche syntaxique.

et que les requêtes utilisées pour l'évaluation contiennent pour la plupart des groupes nominaux. Ces conditions tendent à favoriser un appariement pertinent entre les documents et les requêtes.

3.3.2 Autres méthodes de structuration des termes

Des méthodes différentes ont également été testées. Ainsi, Smeaton [Sme99] propose de recourir à une analyse syntaxique des requêtes et documents pour expliciter les relations entre les unités lexicales à travers une indexation des structures syntaxiques, mais en produisant, dans le cas d'ambiguïtés syntaxiques, des analyses multiples qui sont alors pondérées. Si l'algorithmique assez complexe d'appariement et de pondération mise en œuvre donne de bons résultats en test (appariements syntagme à syntagme), ceux-ci ne sont plus convaincants lors d'une intégration à un SRI. La qualité insuffisante de l'analyseur syntaxique, la complexité de l'algorithme d'appariement et de pondération, les différences entre le style langagier des requêtes et celui des documents[52] et les stratégies de recherche appliquées sont quelques-unes des raisons de cet échec.

Dans un registre un peu différent, Sparck Jones et Tait [ST84] proposent une méthode originale qui consiste à définir des classes d'équivalence de structures propositionnelles, c'est-à-dire de groupes de mots différents mais possédant des constituants identiques et correspondant aux alternatives d'une même relation syntaxique. Les structures [Adjectif + Nom] et [Nom BE Adjectif] sont ainsi membres d'une même classe. Un élément d'une classe, présent dans une requête, peut alors être étendu à l'aide de tous les autres membres.

3.3.3 Utilisation des termes structurés en post-traitement d'un SRI

Il est également utile de faire référence à d'autres expériences qui exploitent les termes structurés non plus au niveau de l'indexation, mais en post-traitement des SRI. Pour illustrer cette idée, nous pouvons citer le système proposé par Mitra *et al.* [MBSC97] qui effectue dans un premier temps un traitement traditionnel des documents et des requêtes, puis procède à une analyse syntaxique des documents retournés (les 100 premiers) par le système en réponse à une question afin d'en extraire les expressions nominales. Une mesure de pondération est ensuite appliquée aux syntagmes extraits. Le système procède alors à une nouvelle indexation de ces documents et crée deux index différents : un pour l'indexation des termes simples et l'autre pour l'indexation des syntagmes. Pour procéder à l'appariement entre la requête et les documents, le système s'appuie sur le score des documents lors de la première interrogation mais prend également en considération, par l'intermédiaire de pondérations spécifiques, les expressions nominales, ce qui permet un reclassement des documents initiaux. Les résultats obtenus ne montrent cependant pas d'amélioration significative, et ne mettent pas suffisamment en valeur l'intérêt de recourir aux expressions nominales en post-traitement d'un SRI. De plus, le temps d'attente pour l'utilisateur est assez important puisque l'analyse syntaxique des documents doit se faire en temps réel.

3.3.4 Bilan de l'apport des termes structurés

À la suite de ces expériences, il semble donc globalement intéressant d'intégrer les termes structurés au sein des SRI. La méthode qui consiste à concevoir des index structurés de type tête+modifieur apparaît efficace puisqu'une amélioration quasi systématique des performances des systèmes est observée. Les expériences menées par Zhai *et al.* [ZTME97][53] confirment cette idée. Leur objectif est de comparer l'impact des différents types d'expressions syntaxiques : les expressions nominales entières (*e.g. heavy construction industry group*), les sous-expressions adjacentes (*e.g. heavy construction industry*), et les paires tête+modifieur (*e.g. industry group, heavy construction, construction industry...*). Bien que les résultats soient dépendants de la façon dont les structures complexes sont intégrées au sein des index et de leur pondération, ils montrent toutefois à nouveau la supériorité de la méthode qui prend en compte les paires tête+modifieur en complément des termes simples, qui conduit à une amélioration de 13% de la précision moyenne et 9% du rappel.

L'influence de la requête ne peut toutefois pas être ignorée (*cf.* les résultats de Strzalkowski *et al.* cités précédemment). Plus une requête est longue, plus elle est descriptive et susceptible de contenir des termes structurés ou des composantes de ces termes.

Les évaluations menées par Smeaton (*cf.* section 3.3.2) quant à l'impact des dépendances syntaxiques peuvent néanmoins contraster avec ces résultats. Il apparaît, dans ses travaux, que cet impact est fortement conditionné par le type et la qualité de l'analyse syntaxique réalisée pour identifier et structurer les syntagmes. Comme le remarque Besançon [Bes02], une analyse syntaxique trop poussée peut parfois conduire à la perte d'informations

52. Un document est généralement composé de phrases bien construites, ce qui n'est pas toujours le cas pour les requêtes.
53. Ces expériences s'inscrivent dans le cadre du projet Claritech NLP.

pourtant utiles pour représenter le contenu de documents [54]. Cette remarque est confirmée par les observations de Hull *et al.* [HGS+97] qui montrent qu'une analyse syntaxique superficielle pour la reconnaissance des expressions, et plus particulièrement des expressions nominales, est suffisante. L'ambiguïté structurelle des termes peut aussi jouer un rôle. À la différence des expressions statistiques qui se limitent généralement à des paires de mots, les syntagmes comportent potentiellement plus d'éléments, ce qui multiplient les risques d'ambiguïtés syntaxiques lors de la segmentation des longues expressions. Une mauvaise décomposition de ces expressions peut être à l'origine de faibles résultats [SLWPC99]. Il est donc parfois nécessaire d'opérer un filtrage de ces syntagmes.

Si les termes complexes et structurés semblent offrir des potentialités aux SRI, il convient néanmoins, avant de conclure à l'intérêt de les prendre en compte, d'évoquer les adaptations à effectuer au cœur de ces systèmes si l'on souhaite exploiter pleinement leur richesse.

3.4 Adaptation des SRI pour l'intégration des informations syntaxiques

Dans cette section, nous nous intéressons plus précisément à certaines spécificités du domaine de la RI, les systèmes développés n'étant pas toujours adaptés à accueillir des informations issues d'analyses linguistiques poussées [55]. Nous avons déjà brièvement évoqué une première difficulté sur laquelle nous revenons ici, en faisant état des solutions expérimentées pour la résoudre : la nécessité de bien pondérer les termes complexes et structurés. Nous nous attachons ensuite à décrire comment ces structures complexes peuvent être intégrées au sein des index, et aux termes simples qu'ils contiennent.

3.4.1 Adaptation des mesures de pondération des termes

Les mesures de pondération utilisées au sein des systèmes doivent être adaptées pour prendre en compte des termes complexes ou structurés. Les mesures traditionnelles (telles que le *term frequency (tf)* $*$ *inverse document frequency (idf))* sont inadaptées à ces termes, moins fréquents que les simples et qui sont alors sous-pondérés. Takenobu *et al.* [THH00] montrent notamment que le facteur *idf* n'est pas adapté aux termes complexes. Or, comme le remarque Sparck Jones [Spa99], une mauvaise pondération de ces structures complexes peut rapidement devenir néfaste au processus de recherche.

À partir de ce constat, un certain nombre de nouvelles mesures de pondération ont été proposées. Une première stratégie consiste à pondérer l'expression (terme complexe ou structuré) en fonction du poids de ses composantes. Les résultats produits ne sont toutefois pas uniformes, comme le prouvent Fagan [Fag87] et Lewis et Croft [LC90]. D'autres auteurs ont recours à une pondération dite « syntaxique ». Haddad [Had02] utilise une pondération de ce type, basée sur les catégories grammaticales des composantes des termes du syntagme. Le poids syntaxique d'une expression est donc calculé en fonction du type de catégories grammaticales de ses composantes [56], et la pondération finale d'indexation est obtenue en multipliant ce poids syntaxique par le poids statistique (*tf*$*$*idf*) de l'expression. Pedersen et Bruce [PB97] utilisent aussi un schéma de pondération basé sur les catégories grammaticales associées aux termes composant une expression, favorisant les syntagmes nominaux ou adjectivaux. Les résultats obtenus montrent une amélioration de précision et rappel. Pour leur système IRENA, Arampatzis *et al.* [ATK96] proposent une mesure de pondération plus sophistiquée des syntagmes qui permet d'accorder davantage d'importance à certains éléments (la tête par exemple).

3.4.2 Stratégies d'intégration des termes complexes ou structurés dans les index

Le deuxième facteur à prendre en compte pour l'exploitation de ces termes en RI est la façon dont ils sont intégrés au sein des index, et leur combinaison avec les termes simples.

Une première méthode consiste à isoler ces deux types d'informations. Ainsi, Haddad [Had03], après avoir évalué différentes stratégies d'intégration au sein du modèle vectoriel, conclut que l'indexation des syntagmes est plus efficace lorsqu'elle est effectuée de manière indépendante de celle des termes simples. Pour cela, suivant une technique initialement proposée par Fox [Fox83], il substitue à une représentation vectorielle classique, regroupant l'ensemble des termes d'indexation au sein d'un seul vecteur, une représentation en deux sous-vecteurs différents. Cette stratégie permet un meilleur classement des documents en tête de liste et accroît la précision. L'utilisation des deux sous-vecteurs est aussi appliquée pour l'expansion des requêtes, qui sont alors enrichies avec des termes provenant de ces deux parties. Kraaij et Pohlmann [KP98] recourent à la même

54. Certaines dépendances entre termes ne correspondant pas à une relation syntaxique (*e.g.* les cooccurrences) peuvent dénoter une information sémantique pertinente.

55. Ces systèmes sont initialement conçus pour prendre en compte uniquement des termes simples.

56. Par exemple, un syntagme dont la tête est un nom aura plus de poids qu'un syntagme dont la tête est un verbe.

technique, mais ces auteurs montrent que les composantes des termes structurés doivent également être ajoutées à l'index des termes simples[57].

Strzalkowski *et al.* [SLWPC99] et Arampatzis *et al.* [AVKV00] choisissent, quant à eux, d'utiliser le même index pour les termes simples et les termes structurés (ou complexes). Ces structures complexes sont utilisées en complément (et non en remplacement) des termes simples, et peuvent également intervenir pour l'expansion des requêtes.

3.5 Conclusion

À l'issue de la description d'un certain nombre d'expériences, la prise en compte des termes complexes ou structurés semble pertinente dans une optique de RI. Plus précisément, l'exploitation de ces structures semble intéressante pour offrir une description plus riche du contenu informationnel. Leur impact sur les performances des systèmes dépend toutefois d'au moins deux facteurs : leur pouvoir de représentativité du contenu textuel – il apparaît que plus ces termes sont significatifs, plus leur exploitation a une influence positive sur les résultats –; la performance des mécanismes de reconnaissance de leurs variantes.

En ce qui concerne plus particulièrement les termes complexes, et bien que les résultats obtenus à travers les diverses expériences évoquées ne permettent pas d'affirmer clairement le type (« statistique » ou « syntaxique ») le plus pertinent dans un cadre de RI, ces deux facteurs vont plutôt dans le sens d'une préférence pour l'exploitation de termes complexes extraits par des méthodes symboliques ou mixtes.

Globalement, il apparaît enfin qu'une analyse syntaxique partielle semble suffisante pour l'extraction et la prise en compte des structures complexes, à condition qu'elle soit capable de gérer les problèmes d'ambiguïté structurelle[58]. Les techniques de TAL paraissent suffisamment souples pour être appliquées au domaine de la RI.

Enfin, il est important de remarquer que l'évaluation et la comparaison de l'impact de ces différents termes sont difficiles à mettre en œuvre. En effet, leur exploitation est fortement dépendante des SRI dans lesquels ils sont intégrés, et plus particulièrement du modèle de représentation utilisé (*e.g.* modèle vectoriel, probabiliste, connexionniste...), des mesures de pondération appliquées, de l'étape du processus de RI au cours de laquelle ils sont intégrés (indexation ou expansion de requêtes)... De même, l'évaluation de leur efficacité sur les résultats des systèmes dépend, comme nous l'avons vu, des conditions d'expérimentations utilisées[59].

57. Au cas où les termes structurés (ou leurs variantes) ne peuvent être utilisés pour établir une correspondance entre les documents et les requêtes, cette stratégie permet tout de même de recourir à leur composante (*i.e.* des termes simples) pour tenter de les apparier.

58. Le cas échéant, il sera nécessaire de procéder à des filtrages pour désambiguïser les structures des syntagmes.

59. Elle est notamment liée au type de collection utilisé (nombre de documents, thème...), aux requêtes (longueur, nombre...), à la langue...

4 Prise en compte de connaissances sémantiques en recherche d'information

Nous terminons notre état de l'art de l'exploitation d'informations linguistiques en RI par la prise en compte de connaissances sémantiques. Nous revenons dans un premier temps sur les divers types de connaissances utilisables, puis nous intéressons à leur intégration dans les SRI, d'une part, lors de la phase d'extension des requêtes et, d'autre part, lors de l'indexation. Un problème récurrent qui se dégage de ces études étant lié au caractère polysémique des informations manipulées, nous terminons cette section par l'évocation de diverses techniques de désambiguïsation automatique utilisés en RI.

4.1 Types d'informations sémantiques exploitables en RI

Avant d'aborder les deux principales façons d'intégrer des informations sémantiques, nous nous penchons sur la provenance de ces connaissances.

4.1.1 Quelles informations sémantiques?

Les bases lexicales existantes sont une des sources d'informations sémantiques exploitables par les SRI. Plusieurs systèmes utilisent ainsi les relations de synonymie ou d'hyperonymie[60] présentes dans la plus connue d'entre elles, WORDNET [Fel98]. Une telle base généraliste n'est toutefois pas forcément adaptée pour un domaine spécialisé ; et la volonté d'utiliser des données plus focalisées se heurte, elle, fréquemment à l'absence de telles ressources. Certains travaux se tournent alors vers l'intégration de connaissances issues d'une acquisition à partir de corpus du domaine d'étude. Pour présenter les informations sémantiques extractibles de ces textes, nous reprenons le découpage lié au type d'approche utilisé pour leur acquisition, déjà énoncé en section 3.1.1 : par une méthode exploitant le caractère soit fréquentiel, soit structurel du corpus[61].

Une première famille de méthodes d'extraction de connaissances sémantiques en corpus se base sur des indices numériques et détecte des associations d'unités lexicales, ou cooccurrences, statistiquement significatives dans des fenêtres de mots ou des contextes syntaxiques. Ces cooccurrences permettent de mettre au jour des termes complexes et des unités en relations syntagmatiques[62] (structures argumentales de verbes par exemple). Rappelons toutefois que les informations sémantiques obtenues par ce seul type de techniques ne sont pas toujours homogènes ou sémantiquement pertinentes. L'approche numérique permet également d'extraire un second type de connaissances sémantiques en analysant les mots qui partagent les mêmes cooccurrents[63] ou plus généralement les mêmes propriétés contextuelles[64]. L'analyse des contextes partagés ou non permet de faire émerger des classes à caractère conceptuel d'unités lexicales et des relations paradigmatiques (synonymie, antonymie, hyperonymie...) les liant. Les informations sémantiques obtenues (classes et relations) ne sont néanmoins pas toujours facilement interprétables[65].

La seconde approche de l'acquisition de connaissances sémantiques (termes complexes, et relations syntagmatiques et paradigmatiques) en corpus exploite des indices structurels – on parle d'approche symbolique. Elle regroupe des travaux reposant sur une expertise linguistique et des études basées sur de l'apprentissage artificiel symbolique [Cla03]. L'expertise linguistique fournit des patrons (morpho-)syntaxiques ou des marqueurs linguistiques (principalement lexicaux) de ce que l'on cherche à repérer en corpus (*cf.* [Jou95] par exemple). Les techniques basées sur l'apprentissage symbolique suivent en général le processus en cinq étapes initié par Hearst [Hea92] : 1- choisir une relation cible $\mathcal{R}$ (*e.g.* la synonymie) ; 2- réunir une liste de paires de mots en relation $\mathcal{R}$ (par exemple les extraire d'un thesaurus, d'une base de connaissances) ; 3- retrouver les phrases du corpus contenant ces paires et enregistrer leurs contextes lexical et syntaxique ; 4- trouver les points communs entre ces contextes et supposer que cela forme un schéma lexico-syntaxique de $\mathcal{R}$; 5- appliquer les schémas pour obtenir de nouvelles paires et retourner en 3. Contrairement aux recherches précédentes, les marqueurs d'une relation sont ici issus d'une analyse d'exemples et non d'une connaissance linguistique *a priori*. La phase 4 peut-être

60. L'hyperonyme est un incluant, un synonyme à un niveau de généralité immédiatement supérieur (*e.g. oiseau* par rapport à *rouge-gorge*).

61. Pour une description plus détaillée, consulter [Cla03].

62. L'approche basée sur la détection de cooccurrences permet d'obtenir des affinités dites de premier ordre [Gre94].

63. Affinités dites du second ordre.

64. Exemples de propriétés contextuelles : contexte syntaxique de mots, relations de dépendance tête+modifieur au sein de syntagmes [BA00], mots cooccurrant dans une certaine fenêtre [PS00]...

65. Les classes regroupent indifféremment des synonymes, antonymes, voire hyperonymes... sans qu'il soit aisé de déterminer le lien précis unissant une unité à une autre.

entièrement manuelle [Hea92, Hea98], ou tirer profit du cadre formel offert par les techniques d'apprentissage symbolique (telles que la programmation logique inductive dans [CSFB03]).

La qualité et l'exploitabilité des informations sémantiques extractibles d'un corpus sont donc liées au genre de méthode d'acquisition choisi. Il est parfois plus difficile de typer précisément le lien unissant des éléments repérés à l'aide de certaines techniques numériques ; et les méthodes symboliques sont en règle générale d'une granularité plus fine mais sont moins portables car elles nécessitent des données initiales (marqueurs, patrons ou exemples) construites manuellement et souvent spécifiques au domaine du corpus[66].

4.1.2 Méthodes d'intégration des informations sémantiques

Comme pour les autres types de connaissances, une première façon d'exploiter des informations sémantiques en RI consiste à utiliser les relations sémantiques disponibles pour étendre les requêtes et accéder à des documents potentiellement pertinents (*via* un synonyme d'un terme de la requête par exemple). L'expansion contribue également à préciser la question de l'utilisateur, puisque l'ajout de termes à la requête initiale cible le sens de ses constituants et les rend moins ambigus. Une seconde façon de procéder est d'introduire des informations sémantiques lors de l'indexation des documents et requêtes afin d'enrichir les représentations. On s'oriente dès lors vers un processus d'indexation plus riche[67], basé non plus sur de simples termes mais sur leur sens, *i.e.* les concepts qu'ils véhiculent, substituant ainsi à l'appariement fondé sur une comparaison de chaînes de caractères un appariement basé sur la signification des termes.

Les informations sémantiques utilisées varient selon le type d'exploitation retenu[68]. Nous choisissons donc de détailler séparément leur intégration pour l'extension de requêtes (section 4.2) et celle dédiée à l'enrichissement de la représentation des documents et requêtes lors de l'indexation (section 4.3).

4.2 Exploitation d'informations sémantiques en extension de requêtes

La nature et les caractéristiques des informations sémantiques extractibles de corpus étant liées à l'approche retenue pour leur acquisition, nous avons choisi de découper ici notre analyse de l'exploitation de ces connaissances en extension de requêtes selon ce critère. Nous nous intéressons tout d'abord, *via* la description de travaux, à l'intégration d'informations extraites à l'aide de méthodes numériques, plus particulièrement de repérage de cooccurrences. Nous nous penchons ensuite sur l'utilisation de relations sémantiques acquises à partir d'une approche symbolique. Nous terminons par l'exploitation d'informations présentes dans des ressources construites *a priori*. Quelques remarques générales concluent cette section.

4.2.1 Utilisation d'informations sémantiques acquises par une approche numérique

Les expérimentations décrites ici enrichissent automatiquement les termes d'une requête d'utilisateur avec les mots qui cooccurrent fortement avec eux dans les documents[69]. Deux stratégies d'extension sont possibles : soit chaque terme de la requête est considéré comme indépendant et est étendu, soit la requête est prise dans sa globalité, et les termes ajoutés doivent être proches de l'ensemble des mots de la requête.

Les expériences de Peat et Willett [PW91] illustrent la première stratégie, sans conduire à une amélioration significative des résultats. Celles, similaires, de Gauch *et al.* [GWR99] montrent également une hausse plutôt faible lors d'expérimentations menées sur la collection Trec-5. Toutefois, leur méthode, appliquée à une collection de documents d'un domaine spécialisé, apporte un accroissement significatif des performances (plus de 28,6% d'amélioration).

Peat et Willett [PW91] justifient globalement la faiblesse ou le caractère mitigé de ces résultats par le fait que les techniques utilisées pour le repérage des cooccurrences favorisent l'extraction de termes de même fréquence. Or, si les mots de la requête sont très fréquents, les termes ajoutés sont alors trop fréquents pour être discriminants (*i.e.* pour permettre de faire une distinction entre documents pertinents et non pertinents). Cette remarque remet donc en cause l'intérêt des méthodes numériques basées essentiellement sur la notion de fréquence.

La seconde stratégie d'expansion possible est de prendre en considération non plus chaque terme de manière isolée, mais la requête dans sa globalité. C'est la stratégie retenue notamment par Qiu et Frei [QF95], qui, au

66. Voir [CS04b] cependant.

67. On parle aussi d'indexation « sémantique ».

68. Il est, en effet, rare de trouver des systèmes qui exploitent les informations sémantiques à la fois pour étendre les requêtes et pour enrichir les index.

69. Par exemple, si les termes *recherche d'information* et *précision* apparaissent souvent ensemble (*i.e.* de façon statistiquement significative) dans les documents, on ajoute *précision* aux requêtes contenant *recherche d'information*.

sein de leurs expériences, réalisent l'expansion de chaque requête non par des termes cooccurrant avec un de ses constituants mais avec le concept qu'elle dénote globalement[70]. Ils montrent que cette approche aboutit à une amélioration de l'efficacité de systèmes comprise entre 20 et 30%.

La particularité de l'expérience de Jing et Croft [JC94] est de ne pas uniquement considérer comme cooccurrents des termes de la requête des termes simples mais également des syntagmes (nominaux, verbaux...). Les cooccurrences syntagmes-termes sont appelées associations, et sont utilisées pour enrichir les requêtes. Trois stratégies principales d'expansion sont mises en œuvre : la première (la *duplication*) a pour objectif de rechercher les syntagmes des documents dont tous les constituants sont des sous-ensembles de la requête. Ces expressions sont ensuite ajoutées aux termes ou syntagmes de la requête. L'objectif est de détecter les expressions importantes de la requête. La seconde stratégie (la *non-duplication*) consiste à repérer, au sein des documents, des syntagmes qui cooccurrent fréquemment avec des termes de la requête. Ces syntagmes sont ajoutés à la requête à condition qu'aucun de leurs termes n'appartienne à la requête. L'idée sous-jacente est de retrouver de nouveaux concepts non exprimés. La dernière stratégie est une combinaison des deux premières méthodes. Les auteurs mesurent également l'impact des différents syntagmes (*i.e.* nominaux, verbaux...). Au final, leurs expériences montrent que l'expansion de requêtes grâce à des cooccurrences entre syntagmes et termes simples améliore les performances (jusqu'à 9,7% d'amélioration), et que ce sont les groupes nominaux sans adjectif qui conduisent aux meilleurs résultats. Cependant, quelques questions demeurent en suspens, telles que le nombre de termes à ajouter à la requête pour obtenir les meilleurs résultats, ou la façon de pondérer ces éléments ajoutés.

4.2.2 Utilisation d'informations sémantiques acquises à l'aide de méthodes symboliques

Parmi les travaux qui emploient des connaissances sémantiques extraites par des techniques symboliques, et obéissant donc à des contraintes linguistiques plus fortes, l'expérience menée par Claveau et Sébillot [CS04a] décrit un enrichissement des noms contenus dans les requêtes par des verbes qui leur sont liés par une relation spécifique (liens dits qualia[71]). L'originalité de la méthode est de considérer que les noms ne sont pas les seules catégories de mots permettant un apport sémantique efficace en reformulation[72]. Les résultats obtenus montrent une amélioration des performances du SRI testé, légère mais statistiquement significative. Il apparaît, en particulier, que les verbes qualia sont utiles pour amener les meilleurs documents en tête de liste.

L'utilisation du lien nom-verbe fait aussi l'objet des travaux de Grefenstette [Gre97]. Partant du principe qu'une des manières de caractériser sémantiquement un nom est d'extraire l'ensemble des verbes utilisés avec lui afin de recenser ce qu'il permet de faire et ce qui est fait en direction de lui (*e.g. show* et *support* pour le nom *research*), le système repère (à l'aide d'une analyse syntaxique partielle des textes et de l'utilisation de patrons[73]) des liens nom-verbe en corpus. En fonction de sa requête, il propose à l'utilisateur de tels liens permettant une définition plus précise de son besoin d'information et contribuant à la désambiguïsation des noms contenus dans la requête. Cette méthode s'avère particulièrement intéressante pour les requêtes courtes, par définition ambiguës.

L'usage de connaissances sémantiques précises acquérables par une approche symbolique semble donc prometteuse. D'autres travaux abondent dans le sens de l'utilité de recourir à des informations linguistiques fines. C'est le cas de ceux de Khoo [Kho95], dont l'idée est d'étendre une requête par des unités lexicales liées par un lien de causalité à ses constituants. Les termes sont étendus soit par des mots partageant la même cause qu'un ou plusieurs d'entre eux (exemple : *cancer du poumon* et *problème respiratoire* partagent la cause *cigarette fumée*), soit par des mots partageant le même effet qu'eux (exemple : *cigarette fumée* et *pollution de l'air* partagent l'effet *cancer du poumon*). Aucune évaluation claire de l'impact de ces liens de causalité n'est cependant fournie[74].

4.2.3 Utilisation d'informations sémantiques contenues dans une base lexicale

Il existe également des études utilisant les connaissances sémantiques contenues dans des bases lexicales existantes. L'exploitation de ces ressources en extension de requêtes consiste à inclure des mots qui sont séman-

70. Ces travaux utilisent une représentation vectorielle des documents et requêtes. Après avoir construit un espace des termes d'indexation (de toute la collection) où la similarité entre chaque paire de termes est calculée, la requête (Q) sous forme d'un vecteur est transposée dans cet espace, et le centroïde de ses termes est calculé ; on obtient alors le concept global « virtuel » de la requête. L'expansion consiste à rechercher dans l'espace des termes ceux qui sont les plus proches de ce concept « virtuel ».

71. L'approche utilisée s'inscrit dans le cadre théorique du Lexique génératif de Pustejovsky [Pus95]. Le repérage de ces liens se fait à l'aide d'un apprentissage artificiel symbolique à partir d'exemples.

72. Les verbes dits qualia ont pour rôle de décrire les différentes facettes sémantiques d'un nom. Il paraît donc intéressant d'utiliser ces verbes en relation avec les noms de la requête pour une reformulation pertinente de cette dernière.

73. Des méthodes statistiques filtrent ensuite les résultats.

74. Le travail est en fait essentiellement centré sur l'identification automatique des relations causales.

tiquement reliés aux concepts de la question originale en suivant les relations de synonymie ou d'hyperonymie qui les structurent très fréquemment. Nous nous focalisons ici uniquement sur les expérimentations mettant en jeu le thesaurus WORDNET [Fel98] qui a souvent été choisi dans un cadre de RI.

Les travaux de Voorhees [Voo98] servent fréquemment de référence pour l'évaluation de l'apport de WORDNET en enrichissement de requêtes. Une expérience, consistant à étendre manuellement des requêtes à l'aide des *synsets*[75] de la base contenant leurs constituants et de hiérarchies liées à ces *synsets*, montre des résultats mitigés, présentant une amélioration essentiellement pour les requêtes courtes. La volonté d'automatisation de l'extension se heurte à la mécanisation extrêmement difficile de la sélection des « bons » *synsets*, *i.e.* des *synsets* correspondant effectivement au sens des termes de la requête[76]. Une autre question récurrente concerne le choix (automatisé) des termes de la requête à étendre.

Un des obstacles majeurs à l'utilisation d'informations sémantiques contenues dans une base lexicale est donc la désambiguïsation automatique. Moldovan et Mihalcea [MM00b] aboutissent aux mêmes conclusions que Voorhees, également en utilisant WORDNET. Smeaton [Sme99], qui expérimente différentes stratégies de pondération des termes ajoutés et diverses tentatives de désambiguïsation automatique, constate également une hausse des performances uniquement dans le cas d'un choix manuel des *synsets*.

4.2.4 Bilan

Si l'on cherche à faire un bilan de ce qui vient d'être présenté, on constate que d'une manière générale, l'enrichissement des requêtes est intéressant uniquement si les mots ajoutés sont véritablement liés aux mots de la requête. Tout l'art de cette technique réside donc dans la reconnaissance du sens des termes de la question. L'utilisation de ressources généralistes telles que WORDNET présente le danger de ne pas contenir le sens précis des mots de la requête, ou de proposer trop de sens différents ; elle doit donc nécessairement être accompagnée d'un traitement de désambiguïsation efficace.

Les expériences décrites ont également soulevé plusieurs problèmes récurrents. Le premier, comme le souligne Voorhees, est de déterminer les termes à étendre et leur sélection automatique. Une solution ou alternative est d'adopter la démarche proposée par Qiu et Frei [QF95] consistant à identifier le concept global de la requête, ce qui n'est possible que si tous les termes choisis par l'utilisateur pour formuler son besoin d'information font référence au même concept. Une autre stratégie est de recourir, pour l'expansion de requêtes, à la rétroaction de pertinence (*relevance feedback* en Anglais)[77]. Son principe général est d'exploiter les documents retournés en réponse à une requête et de les utiliser pour améliorer la qualité des résultats. L'approche de Xu et Croft [XC00] va dans ce sens. Après extraction, dans les documents retournés initialement par le système, des groupes nominaux (appelés concepts), et classement de ceux-ci en fonction de leurs cooccurrences avec un des termes de la requête, les groupes les mieux classés sont utilisés en expansion. Les résultats obtenus sont plutôt positifs[78], à condition cependant que les documents retournés la première fois soient pertinents.

De plus, comme l'attestent les travaux de Gauch *et al.* [GWR99], l'expansion de requêtes apparaît particulièrement efficace sur des corpus spécialisés. Une solution pour étendre cette efficacité à des domaines plus ouverts serait d'identifier le domaine de connaissance auquel fait référence la requête, afin de pouvoir l'enrichir avec des mots du même domaine et sémantiquement similaires. Au sein d'un domaine de connaissance, la polysémie d'un terme étant réduite, l'expansion pourrait alors être plus précise[79]. Cependant cette tâche est difficile.

Pour pallier les limites des diverses approches de l'acquisition d'informations sémantiques évoquées précédemment, Mandala *et al.* [MTT99] proposent, pour leur part, une technique d'enrichissement des requêtes basée sur la combinaison de connaissances extraites à l'aide de plusieurs méthodes, couplée à l'utilisation de mesures de pondération des termes de la requête à étendre. Une première stratégie choisie[80] consiste à utiliser uniquement WORDNET ; la deuxième s'appuie sur des relations tête+modifieur obtenues par une approche symbolique ; la troisième exploite les informations sémantiques acquises par repérage de cooccurrences. La dernière mèle l'ensemble de ces connaissances pour enrichir les requêtes. Les résultats obtenus montrent une amélioration significative des performances[81] lors du couplage des différentes informations sémantiques[82]. Cette méthode

75. Classes de synonymes dans WORDNET.

76. Le nombre de *synsets* auxquels appartient un mot dépend du nombre de ses sens.

77. Telle qu'elle a été définie par Rocchio [Roc71].

78. Amélioration de la précision moyenne de 14,3% pour les requêtes longues et de 24,9% pour des requêtes moyennes (environ 7 termes).

79. Il serait possible par exemple, d'utiliser des ressources lexicales qui recensent les différents domaines auxquels un terme peut être attaché, et de solliciter l'utilisateur pour sélectionner le domaine pertinent.

80. Expériences menées sur la collection de test Trec-7.

81. Une amélioration de 39% de la précision moyenne.

82. La méthode de pondération des termes à étendre joue également un rôle important. Pour plus de précision, *cf.* [MTT99].

présente l'intérêt de désambiguïser une partie des mots de la requête et d'éviter ainsi leur enrichissement par des termes inadéquats.

4.3 Exploitation d'informations sémantiques pour l'indexation

Même si la frontière est parfois un peu diffuse, on distingue principalement deux types d'exploitation de connaissances sémantiques pour l'indexation, que nous présentons successivement ici. Certains SRI optent pour une indexation conceptuelle, généralement fondée sur des ontologies, applicable sur des domaines spécialisés et faisant usage d'un formalisme de représentation de connaissances. D'autres sont basés sur une indexation sémantique, qui utilise, pour enrichir l'indexation des documents et requêtes, des informations sémantiques soit généralistes, soit acquises en corpus, et est accompagnée d'un traitement de l'ambiguïté.

4.3.1 Indexation conceptuelle

Le principe de fonctionnement des systèmes basés sur une indexation conceptuelle est le suivant : à l'aide de bases de connaissances lexicales[83] du domaine structurées[84], de techniques de TAL, voire d'apprentissage pour acquérir des connaissances nouvelles sur le domaine analysé, les termes significatifs et les liens qui les unissent sont extraits des documents et requêtes, et sont décrits à l'aide d'un formalisme de représentation de connaissances (graphes conceptuels, réseaux sémantiques[85], logiques de description[86]...) qui permet d'identifier les concepts et d'expliciter leurs relations sémantiques.

Woods et Ambroziak [WA98] proposent ainsi un exemple d'indexation conceptuelle basé sur une organisation taxonomique de la connaissance. Une taxonomie conceptuelle organisée selon une relation de subsomption (*i.e.* reliant des concepts plus généraux aux plus spécifiques qu'ils subsument) est utilisée pour tenter des appariements entre les termes de la requête et les documents[87]. Cette taxonomie est construite à partir des mots et expressions extraits des textes, grâce à une base lexicale existante, où sont représentées des relations de subsomption généralistes de la langue, et d'une analyse morphologique et syntaxique des textes[88]. Sur de petites collections de documents techniques, les auteurs montrent que l'indexation conceptuelle améliore la précision et le rappel par rapport à un système d'indexation classique.

Le projet ONTOSEEK développé par Guarino *et al.* [GMV99] est un système de recherche documentaire développé pour des documents de type catalogue (pages jaunes, catalogues commerciaux...) dont la structure est pratiquement identique. Sa particularité est de coupler une représentation des connaissances du domaine basée sur un formalisme variante des graphes conceptuels de Sowa [Sow84] (les *lexical conceptual graphs*) à une ontologie générale (WORDNET). Les documents sont représentés sous forme de graphes conceptuels[89]. La requête est représentée sous la même forme[90]. La recherche de documents consiste alors en un appariement de graphes dirigé par les ontologies, en se basant notamment sur la relation de subsomption.

ELEN (géniE logicieL recherchE d'informatioNs), mis en œuvre par Chevallet [Che92], est un prototype de système conceptuel de RI, destiné à l'interrogation de logiciels et de leur documentation associée, qui repose sur une représentation des connaissances par graphes conceptuels de Sowa [Sow84]. Une requête, d'abord exprimée dans un pseudo-langage graphique par l'utilisateur, est traduite en un graphe conceptuel. La recherche, quant à elle, est fondée sur l'opération de projection (*i.e.* du graphe de la requête sur les graphes représentant les documents).

Le projet MENELAS [ZM94], dédié au domaine médical, est également basé sur les graphes conceptuels. Il réalise une indexation automatique par le biais d'une grammaire contextuelle et d'une analyse sophistiquée, et utilise un treillis d'environ 1000 concepts. Les requêtes sont formulées grâce à une interface graphique qui oriente l'utilisateur vers une requête non ambiguë.

RIME (Recherche d'Informations MÉdicales) [Ber90] est un SRI, également spécialisé dans le domaine médical, basé sur un modèle sémantique s'inspirant des dépendances conceptuelles de Schank [Sch72]. Il utilise des représentations arborescentes dans lesquelles les feuilles sont des concepts primitifs du domaine et les nœuds

83. Contenant des unités lexicales et des informations sur leurs propriétés (d'ordre morphologique, sémantique...).
84. Par des relations telles que *est-un* (*is-a* en Anglais)...
85. Généralement représentés par des graphes orientés dont les sommets sont étiquetés par des concepts et les arcs par des relations entre concepts.
86. Fournissant des mécanismes d'inférence comme la subsomption.
87. L'exemple donné par Woods et Ambroziak est que leur système peut automatiquement déterminer que *car washing* est un type de *automobile cleaning* s'il possède l'information que *car* est une sorte d'*automobile* et que *washing* est un type de *cleaning*.
88. Ces analyses permettent de gérer le problème des variantes morphologiques et de prendre en compte les termes complexes.
89. Le système assiste cette phase, en se basant sur WORDNET et sur un certain nombre de règles.
90. Une assistance interactive est utilisée pour la formulation de la requête.

non primitifs des opérateurs sémantiques[91]. Il réalise une mise en correspondance documents-requête reposant sur le modèle logique de Van Rijsbergen [Van91], et s'appuyant sur la recherche d'une implication logique entre la requête et le contenu des documents par rapport à un ensemble de connaissances pré-établies. Ce SRI emploie un vocabulaire très spécialisé et s'appuie également sur une grammaire contextuelle spécifique[92].

L'indexation conceptuelle peut représenter une solution à certains problèmes du langage naturel (tels que la polysémie par exemple) en permettant une indexation non plus à base de mots mais fondée sur la notion de concepts et sur les relations entre ces concepts. Elle nécessite cependant des ressources considérables (*i.e.* des bases de connaissances de domaines), et fait appel à des techniques complexes (*i.e.* généralement issues de l'intelligence artificielle)[93]. Un de ses points faibles majeurs est donc lié au coût de sa mise en œuvre. De plus, les performances de ces SRI dit « conceptuels » n'ont pas atteint un stade supérieur à celles des SRI traditionnels. Ces remarques nous amènent à considérer un autre type d'indexation, l'indexation sémantique, plus souple que la précédente.

4.3.2 Indexation sémantique

Les SRI basés sur une indexation sémantique utilisent des ressources soit construites manuellement, soit acquises en corpus, pour enrichir la représentation des documents et requêtes à l'aide de mots sémantiquement proches de ceux extraits de leur contenu.

Utilisation de ressources *a priori*

La base lexicale WORDNET est d'usage fréquent dans de tels SRI. Après avoir appliqué un traitement de désambiguïsation aux termes des documents et requêtes, l'indexation est enrichie par l'ajout des *synsets* leur correspondant dans la base.

Les expériences de Mihalcea et Moldovan [MM00a] reposent sur ce principe, et présentent la particularité supplémentaire de prendre en compte les termes complexes, indexés indépendamment des simples. Les résultats obtenus montrent que le couplage indexation à base de mots-clés - indexation à base de *synsets* permet une augmentation de 16% du rappel et de 4% de la précision. Les auteurs soulignent toutefois qu'un traitement plus efficace de la désambiguïsation[94] produirait des performances encore plus élevées, ce qui rappelle un principe que nous avons évoqué en section 4.2.3.

Smeaton et Quigley [SQ96] proposent également une expérimentation d'indexation[95] fondée sur les *synsets*, qui conduit à une hausse de 29% de la précision, mais dans laquelle la désambiguïsation des termes est manuelle. C'est aussi le cas dans les travaux de Gonzalo *et al.* [GVCC98] qui montrent un accroissement identique de 29% de la précision par rapport à une indexation traditionnelle. Ces mêmes auteurs tentent par ailleurs une expérience dans laquelle seule la désambiguïsation manuelle des termes de la requête est effectuée. L'indexation à l'aide des *synsets* de WORDNET améliore aussi les performances, les documents contenant à la fois les termes de la requête et des membres des *synsets* associés étant plus pertinents que des textes contenant uniquement les mots initiaux. Ceci amène donc à envisager une intervention de l'utilisateur pour désambiguïser les termes de sa requête, en particulier quand celle-ci est courte et se prête mal (contexte insuffisant) à un processus de désambiguïsation automatique.

Utilisation d'informations sémantiques acquises en corpus

Les SRI à indexation sémantique basée sur des connaissances acquises en corpus exploitent généralement des informations de cooccurrences et de similarités de cooccurrents pour dériver les sens des termes et aboutir à un appariement fondé sur ces sens et non plus sur les seuls mots.

Ainsi Schütze et Pedersen [SP95] proposent une approche dans laquelle documents et requêtes sont tout d'abord traités pour extraire les termes d'indexation et leurs sens. Les documents sont alors classés en fonction du nombre de sens qu'ils partagent avec la requête. L'attribution d'un sens à chaque terme repose sur l'exploration de son contexte et sur le principe que les occurrences d'un mot utilisées dans le même sens partagent les mêmes contextes. Pour chaque terme, un algorithme calcule ses vecteurs contextuels, *i.e.* les mots voisins de chacune

91. Qui permettent l'expression de concepts structurés en combinant soit les concepts primitifs, soit les concepts structurés déjà construits.

92. Dans le domaine très spécialisé du langage médical utilisé, il est possible d'établir une représentation interne correspondant à la sémantique d'une phrase.

93. Une complexité qui est également liée au niveau de granularité souhaitée pour l'application.

94. L'algorithme actuel de désambiguïsation automatique effectue un certain nombre de traitements itératifs (identification des noms propres, identification des mots qui ont un seul sens dans WORDNET ; pour un mot w_i, identification des paires (w_{i-1}, w_i) et (w_i, w_{i+1}) et extraction des occurrences de ces paires pour appliquer un certain nombre d'heuristiques...).

95. L'expérience consiste à indexer de très courts documents (titres d'images).

de ses occurrences. Les vecteurs contextuels d'un mot sont partionnés en régions (une par sens), un vecteur étant affecté à une région s'il est plus proche du centroïde du *cluster* de cette région que des autres centroïdes. Chaque occurrence d'un terme est donc classée comme ayant un sens ou un autre, et ce traitement est effectué pour tous les termes des documents. Les informations en sortie sont stockées au sein d'une structure appelée thesaurus. Les termes de la requête subissent le même traitement : calcul des vecteurs contextuels, assignation de sens (celui du centroïde de *cluster* le plus proche). Les résultats montrent une augmentation de 4% de la précision moyenne. La fusion de la liste ordonnée de documents produite avec ce type d'indexation et de celle obtenue avec une indexation classique améliore la précision de 11%.

L'hypothèse sous-jacente aux travaux de Rajman *et al.* [RBC00] dans leur système DSIR est qu'il est possible d'améliorer les performances d'un SRI (vectoriel, dans leur cas) en ajoutant des connaissances sémantiques, qui sont des informations de cooccurrence d'unités linguistiques (noms, verbes...) des documents avec les termes d'indexation retenus. La probabilité d'associer le terme t_j à un document est alors calculée à partir de la fréquence des unités linguistiques dans le document et du nombre de cooccurrences de ces unités avec t_j. Ce nombre de cooccurrences d'un mot avec le terme d'indexation t_j est vu comme un estimateur de la probabilité que ce mot exprime le sens t_j. Les documents restent donc représentés dans un espace vectoriel de dimension égale à la taille de l'ensemble des termes d'indexation, bien que le modèle tienne compte d'un plus grand nombre d'unités linguistiques. Le SRI testé présente une amélioration très significative des performances pour un faible rappel, moins au-delà.

En conclusion, signalons que les méthodes d'indexation qui proposent d'exploiter le sens des mots sont encore peu utilisées dans les SRI. La dernière méthode proposée, qui extrait la connaissance utile de corpus et semble capable de s'adapter à différentes collections, donne des pistes intéressantes mais encore à explorer plus en profondeur. Une telle indexation sémantique est toutefois tributaire d'une désambiguïsation efficace des termes, qui doit forcément, dans un cadre de RI, être automatique. Nous nous intéressons donc, pour terminer cette section, aux méthodes de désambiguïsation qui ont été appliquées à la RI.

4.4 Désambiguïsation et RI

La désambiguïsation automatique est un champ de recherche à part entière, qui pourrait à lui seul faire l'objet d'un état de l'art. Ce n'est pas le but de ce document et le lecteur intéressé pourra se référer à [San97, CHU00] ou [Aud03] par exemple. Même dans le cadre de la RI, de nombreuses recherches ont eu lieu, parmi lesquelles nous allons nous contenter ici de présenter quelques méthodes et expériences.

On peut globalement distinguer deux approches principales de la désambiguïsation, en particulier en RI. La première est basée sur une connaissance *a priori* du nombre de sens d'un mot et repose généralement sur l'utilisation de ressources sémantiques telles que des dictionnaires, où les mots sont représentés par des définitions, ou bien des thesaurus ou bases lexicales sémantiques telles que WORDNET par exemple. Il s'agit alors de reconnaître le sens d'une occurrence donnée d'un mot parmi ses diverses significations possibles. La seconde s'appuie sur des informations extraites directement des documents manipulés, à l'aide principalement de méthodes numériques ; la désambiguïsation consiste dans ce cas à déterminer si un mot ambigu donné est utilisé avec le même sens dans deux occurrences distinctes, plutôt que de chercher à associer un sens précis à ce mot. La connaissance *a priori* du nombre de sens possibles n'est pas indispensable.

Parmi les systèmes à base de dictionnaires, Wallis [Wal93] propose une technique de désambiguïsation (fondée sur les travaux de Wilks *et al.* [WFG+90]) qui remplace les mots contenus dans les documents par leurs définitions issues du dictionnaire LDOCE[96], et part du principe que deux mots synonymes auront des définitions similaires. Son introduction au sein de SRI ne permet toutefois pas de conclure à un apport effectif[97].

Voorhees [Voo98] utilise quant à elle WORDNET, et associe un *synset* à un mot donné en fonction du nombre de correspondances entre le contexte du mot et le contexte d'un *synset*[98] du mot dans la hiérarchie de WORDNET. Le sens d'un mot ambigu (*i.e.* appartenant donc à plusieurs *synsets*) dans une phrase donnée est déterminé en calculant la distance sémantique entre chaque *synset* de ce mot et ceux des autres termes de la phrase ; le *synset* le plus proche des autres mots de la phrase est alors choisi. L'hypothèse sous-jacente est que le sens approprié d'un mot ambigu doit être proche des sens des autres mots de son contexte. Les expérimentations se révèlent décevantes puisque l'intégration de cette méthode de désambiguïsation dans un SRI dégrade les performances,

96. Longman Dictionary of Contemporary English.

97. Expériences réalisées sur les collections de tests Cacm et Time.

98. Le contexte d'un synset s est défini comme « le plus grand sous-graphe connexe contenant s, contenant uniquement des descendants d'un ancêtre de s et ne contenant pas de *synset* incluant un descendant qui comprend une autre instance d'un membre de s ».

ce qui peut en partie s'expliquer par la difficulté de désambiguïser les mots contenus dans les requêtes courtes (composées d'un ou deux mots).

Les travaux de Uzuner *et al.* [UKY99] mettent aussi en œuvre un système de désambiguïsation basé sur l'utilisation de *synsets* et se focalisent sur le contexte local des mots ambigus. L'idée-clé est que les mots utilisés avec le même contexte (les unités lexicales de ce contexte sont appelées sélecteurs) ont des sens similaires ou reliés. Les informations contenues dans WORDNET combinées aux sélecteurs permettent d'identifier le *synset* approprié du mot ambigu dans son contexte. L'évaluation du système sur le corpus SemCor révèle un taux de précision de 60%, mais son intégration dans un SRI (SMART) ne montre pas d'amélioration de résultats. Les erreurs de désambiguïsation excluent des synonymes corrects qui auraient dû être utilisés pour étendre des requêtes, et introduisent des informations incorrectes entraînant une dégradation des performances.

Des résultats plus positifs sont obtenus par Moldovan et Mihalcea [MM00b] qui proposent une expérience de désambiguïsation basée sur des informations contextuelles et WORDNET. L'intérêt du système de désambiguïsation mis en œuvre réside dans le fait qu'il détermine au préalable les mots qui pourront être désambiguïsés[99] avec une bonne précision, les autres mots n'étant pas traités. En procédant ainsi, il s'assure que la désambiguïsation n'est pas destructrice, tout en étant rapide et robuste (55% des noms et des verbes sont désambiguïsés avec une précision de 92,2%).

Les résultats globalement mitigés de désambiguïsation obtenus avec WORDNET ont diverses causes, parmi lesquelles on peut citer sa couverture insuffisante ou la granularité trop fine des sens, difficilement compatible avec la RI[100]...

L'apprentissage supervisé appliqué à des corpus sémantiquement étiquetés à la main[101] est une autre façon de réaliser une désambiguïsation automatique (*cf.* par exemple les travaux de Yarowski [Yar95, Yar00], Towell et Voorhees [TV98] ou Ng et Lee [NL02]). Toutefois les études utilisant des méthodes de ce type ont généralement besoin de corpus d'entraînement volumineux[102], peu courants actuellement.

Les approches proposées par Schütze et Pedersen [SP97] ou Rajman *et al.* [RBC00] décrites plus haut s'appuient, quant à elles, sur le seul contexte des mots pour opérer la désambiguïsation ; en particulier dans la première, les termes sont regroupés selon les similitudes de leurs contextes, et chaque groupe est supposé être représentatif d'un sens. Les sens déterminés ainsi sont très liés au corpus sur lequel ils ont été appris et sont peu réutilisables, mais le principal inconvénient des techniques proposées est qu'elles sont assez mal adaptées à la désambiguïsation de termes contenus dans des requêtes courtes.

En conclusion, il apparaît difficile d'évaluer la méthode de désambiguïsation la plus efficace dans une optique de RI. L'utilisation de ressources construites *a priori* paraît limitée pour plusieurs raisons : les sens des mots y sont généralement statiques et ne sont pas forcément adaptés à un corpus donné, éventuellement spécialisé. Les informations issues des corpus semblent potentiellement plus prometteuses, mais leur intégration dans des SRI produit actuellement des résultats qui ne sont pas totalement convaincants.

4.5 Conclusion

L'efficacité de l'exploitation de connaissances sémantiques en RI dépend donc d'un nombre assez conséquent de facteurs. Un d'entre eux est leur provenance ou leur mode d'acquisition, acquisition dont la complexité doit d'ailleurs aussi être prise en compte. Les diverses approches expérimentées (*i.e.* utilisant des informations issues de ressources généralistes construites *a priori*, ou extraites à l'aide de méthodes numériques ou symboliques) possèdent toutes des inconvénients. Une solution envisageable pour pallier leurs limites respectives et exploiter au mieux les connaissances sémantiques serait de combiner des informations obtenues par ces différents modes. Il convient également de garder à l'esprit le problème de l'adaptabilité des informations extraites à l'évolution des collections traitées dans un cadre de RI.

S'interroger sur la façon la plus efficace d'intégrer ces connaissances sémantiques au sein des systèmes est aussi un point essentiel : *a posteriori* du processus d'indexation, c'est-à-dire en extension de requêtes, ou bien en amont, au cœur même du système. Le second choix nécessite d'adapter les modèles de RI, traditionnellement non prévus pour prendre en compte des relations de dépendances entre termes.

99. Pour cela, le système s'appuie sur un traitement de désambiguïsation (décrit en note 94) itératif qui s'interrompt lorsque plus aucun terme ne peut être désambiguïsé. Les mots ambigus restants sont alors ignorés.
100. Ainsi, WORDNET répertorie pour le nom *break* plus de 63 sens différents.
101. Un exemple de corpus étiqueté sémantiquement est le corpus SemCor [Fel98] qui contient 250000 mots étiquetés.
102. Pour plus de précision sur ces techniques, se référer à [Aud03].

5 Conclusion

La synthèse que nous venons de présenter des contributions possibles des techniques du TAL à la RI a montré que ces techniques, à travers une analyse plus fine des documents et requêtes prenant en compte différents niveaux de la langue, permettent d'extraire des informations plus riches que de simples mots-clés. Ces connaissances favorisent une meilleure représentation du contenu informationnel et du besoin de l'utilisateur. Ayant fait le choix de dresser en fin de chaque section un bilan partiel de l'apport des divers types de données, nous nous contentons de souligner ici quelques pistes encore à explorer dans ce vaste domaine, où il convient de trouver un équilibre entre la finesse de la description linguistique et son exploitation efficace.

Une première constatation que l'on peut faire, qui ouvre de nombreuses perspectives de travaux, est que les potentialités du TAL sont encore loin d'être totalement exploitées. Sur le plan syntaxique, par exemple, seuls sont actuellement pris en compte les termes complexes et structurés, alors que d'autres connaissances structurelles et structurantes telles que les entités nommées (noms de lieux, de personnes, de sociétés...) mais également des informations de positionnement des mots (distance, ordre...) sont extractibles. À un niveau plus sémantique, rares sont les expériences qui s'intéressent à la notion de thème, c'est-à-dire qui cherchent à identifier le ou les thèmes des documents et à établir une correspondance avec la thématique des informations recherchées par l'utilisateur. Les travaux en détection et caractérisation des sujets abordés ont pourtant atteint un niveau assez intéressant [RS03, FG02] pour être testés dans ce cadre, et pour chercher à tirer profit de la connaissance du thème d'un document afin de limiter les problèmes d'ambiguïté.

Par ailleurs, les différentes expérimentations effectuées jusqu'à présent ne prennent généralement en compte que des informations linguistiques d'un seul niveau de langue (morphologique, syntaxique ou sémantique). Il serait intéressant de chercher à coupler ces diverses connaissances pour voir si une telle caractérisation très riche des documents et requêtes a effectivement une influence positive en RI. Une façon d'intégrer conjointement ces informations linguistiques pourrait être d'utiliser l'architecture proposée par Strzalkowski *et al.* [SLWPC99] basée sur un système d'index parallèles, où chaque index reflète une stratégie particulière de représentation des textes[103]. Au final, ce « méta-système » de recherche peut être optimisé en utilisant les meilleurs résultats de chaque stratégie. Toujours sur le plan des couplages, développer et évaluer des systèmes exploitant à la fois l'extension des requêtes et l'indexation riche des documents et questions est aussi une piste à explorer.

Il convient toutefois d'avoir à l'esprit que la palette d'informations linguistiques effectivement exploitables est en fait très dépendante des modèles de RI dans lesquels elles vont être intégrées. Actuellement, la plupart des SRI se fondent sur des modèles de représentation des documents et requêtes dont la principale caractéristique est de reposer sur des ensembles de mots indépendants. Il est évident que ce type de représentation n'est pas adapté à des informations qui cherchent à établir des relations entre les termes. Les capacités de ces modèles à intégrer de telles connaissances est donc encore à étudier, et leurs limites actuelles à la perméabilité à ces ressources est à mettre au jour. Ceci conduira éventuellement à la conception de nouveaux modèles de représentation plus aptes à exploiter pleinement la puissance des informations linguistiques.

103. On pourrait ainsi avoir un index qui prendrait en compte les informations sémantiques, un autre pour les informations syntaxiques...

Références

[ATK96] A. Arampatzis, T. Tsoris, and C. Koster. IRENA: Information Retrieval Engine Based on Natural Language Analysis. Technical Report CSI-R9623, Computing Science Institute, Université de Nijmegen, Pays-Bas, 1996.

[Aud03] L. Audibert. *Outils d'exploration de corpus et désambiguïsation lexicale automatique.* PhD thesis, Université d'Aix-Marseille I - Université de Provence, Aix-en-Provence, France, 2003.

[AVKV00] A. Arampatzis, T.P. Van Der Weide, C.H.A. Koster, and P. Van Bommel. *Linguistically Motivated Information Retrieval*, volume 69, pages 201–222. M. Dekker, New York, États-Unis, 2000.

[BA00] D. Bourigault and H. Assadi. Analyse syntaxique et analyse statistique pour la construction d'ontologie à partir de textes. In D. Bourigault, J. Charlet, G. Kassel, and M. Zacklad, editors, *Ingénierie des connaissances. Tendances actuelles et nouveaux défis*, pages 243–255. Éditions Eyrolles/France Télécom, Paris, France, 2000.

[Ber90] C. Berrut. Indexing Medical Reports - The RIME Approach. *Information Processing and Management*, 26(1):93–109, 1990.

[Bes02] R. Besançon. *Intégration de connaissances syntaxiques et sémantiques dans les représentations vectorielles de textes. Application au calcul de similarités sémantiques dans le cadre du modèle* DSIR. PhD thesis, École Polytechnique Fédérale de Lausanne, Suisse, 2002.

[CDHK01] J. Carlberger, H. Dalianis, M. Hassel, and O. Knutsson. Improving Precision in Information Retrieval for Swedish Using Stemming. In *Proceedings of the 13th Nordic Conference on Computational Linguistics, NODALIDA 01*, Uppsala, Suède, 2001.

[CH01] J.-P. Chevallet and M.H. Haddad. Proposition d'un modèle relationnel d'indexation syntagmatique : mise en œuvre dans le système IOTA. In *Proceedings of 19e Congrès Informatique des organisations et systèmes d'information et de décision, INFORSID 2001*, Martigny, Suisse, 2001.

[Che92] J.-P. Chevallet. *Un modèle logique de recherche d'information appliqué au formalisme des graphes conceptuels. Le prototype* ELEN *et son expérimentation sur un corpus de composants logiciels.* PhD thesis, Université Joseph Fourier, Grenoble, France, 1992.

[Chu95] K.W. Church. One Term or two? In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, États-Unis, 1995.

[CHU00] Special Issue on Senseval. *Computers and the Humanities*, 34(1/2), 2000.

[Cla03] V. Claveau. *Acquisition automatique de lexiques sémantiques pour la recherche d'information.* PhD thesis, Université de Rennes 1, France, 2003.

[CS04a] V. Claveau and P. Sébillot. Extension de requêtes par lien sémantique nom-verbe acquis sur corpus. In *Proceedings of 11e conférence de Traitement automatique des langues naturelles, TALN'04*, Fès, Maroc, 2004.

[CS04b] V. Claveau and P. Sébillot. From Efficiency to Portability: Acquisition of Semantic Relations by Semi-Supervised Machine Learning. In *Proceedings of the 20th International Conference on Computational Linguistics, COLING'04*, Genève, Suisse, 2004.

[CSFB03] V. Claveau, P. Sébillot, C. Fabre, and P. Bouillon. Learning Semantic Lexicons from a Part-of-Speech and Semantically Tagged Corpus Using Inductive Logic Programming. *Journal of Machine Learning Research, Special Issue on Inductive Logic Programming*, 4:493–525, 2003.

[Dai96] B. Daille. Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In P. Resnik and J. Klavans, editors, *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, pages 49–66. The MIT Press, 1996.

[Dai02] B. Daille. *Découvertes linguistiques en corpus.* Mémoire d'habilitation à diriger des recherches, Université de Nantes, France, 2002.

[DFS02] B. Daille, C. Fabre, and P. Sébillot. Applications of Computational Morphology. In P. Boucher, editor, *Many Morphologies*, pages 210–234. Cascadilla Press, Somerville, 2002.

[DG83] G.M. Dillon and A.S. Gray. FASIT: A Fully Automatic Syntactically Based Indexing System. *Journal of the American Society for Information Science*, 34(2):99–108, 1983.

[DN00] G. Dal and F. Namer. Génération et analyse automatiques de ressources lexicales construites utilisables en recherche d'informations. *Traitement Automatique des Langues*, 41(2):423–446, 2000.

[Fag87] J. Fagan. *Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Non-Syntactic Methods.* PhD thesis, Université de Cornell, New-York, États-Unis, 1987.

[Fel98] C. Fellbaum, editor. WORDNET*: An Electronic Lexical Database.* The MIT Press, Cambridge, Massachussets, États-Unis, 1998.

[FG02] O. Ferret and B. Grau. A Bootstrapping Approach for Robust Topic Analysis. *Natural Language Engineering (NLE), Special issue on robust methods of corpus analysis*, 8(3):209–233, 2002.

[FGM+96] N. Faraj, R. Godin, R. Missaoui, S. David, and P. Plante. Analyse d'une méthode d'indexation automatique basée sur une analyse syntaxique de texte. *Canadian Journal of Information and Library Science / Revue canadienne des sciences de l'information et de bibliothéconomie*, 21(1):1–21, 1996.

[Fox83] E.A. Fox. *Extending the Boolean and Vector Space Models of Information Retrieval with P-norm Queries and Multiple Concept Types.* PhD thesis, Cornell University, New-York, États-Unis, 1983.

[FZ98] M. Fuller and J. Zobel. Conflation-Based Comparison of Stemming Algorithms. In *Proceedings of the 3th Australian Document Computing Symposium*, Sydney, Australie, 1998.

[GGHR00] É. Gaussier, G. Grefenstette, D. Hull, and R. Roux. Recherche d'information en Français et traitement automatique des langues. *Traitement automatique des langues*, 41(2):473–493, 2000.

[GGS97] É. Gaussier, G. Grefenstette, and M. Schulze. Traitement du langage naturel et recherche d'informations : quelques expériences sur le Français. In *Proceedings of 1es journées scientifiques et techniques du Réseau francophone de l'Ingénierie de la langue de l'AUPELF-UREF*, Avignon, France, 1997.

[GMV99] N. Guarino, C. Massolo, and G. Vetere. ONTOSEEK: Content-Based Access to the Web. *IEEE Intelligent Systems*, 14(3):70–80, 1999.

[Gre94] G. Grefenstette. Corpus-Derived First, Second and Third-Order Word Affinities. In *Proceedings of EURALEX International Congress*, Amsterdam, Pays-Bas, 1994.

[Gre97] G. Grefenstette. SQLET: Short Query Linguistic Expansion Techniques: Palliating One-Word Queries by Providing Intermediate Structure to Text. In *Proceedings of the 5th International Conference Recherche d'Informations Assistée par Ordinateur, RIAO 97*, Montréal, Canada, 1997.

[Gro96] G. Gross. *Les expressions figées en Français : noms composés et autres locutions.* Ophrys, Paris, France, 1996.

[GVCC98] J. Gonzalo, F. Verdejo, I. Chugur, and I.J. Cigarran. Indexing with WORDNET Synsets can Improve Text Retrieval. In *Proceedings of the COLING/ACL '98 Workshop on Usage of* WORDNET *for NLP*, Montréal, Canada, 1998.

[GWR99] S. Gauch, J. Wang, and S.M. Rachakonda. A Corpus Analysis Approach for Automatic Query Expansion and its Extension to Multiple Databases. *ACM Transactions on Information Systems*, 17(3):250–269, 1999.

[Had02] H. Haddad. *Extraction et impact des connaissances sur les performances des systèmes de recherche d'information.* PhD thesis, Université Joseph Fourier, Grenoble, France, 2002.

[Had03] H. Haddad. Utilisation des syntagmes nominaux dans un système de recherche d'information. In *Proceedings of 19es Journées de Bases de Données Avancées, BDA*, Lyon, France, 2003.

[Har91] D. Harman. How Effective is Suffixing? *Journal of the American Society for Information Science*, 42(1):7–15, 1991.

[Hea92] M.A. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the 14th International Conference on Computational Linguistics, COLING'92*, Nantes, France, 1992.

[Hea98] M.A. Hearst. Automated Discovery of WordNet Relations. In Christiane Fellbaum, editor, *WordNet: an Electronic Lexical Database*, chapter 5, pages 131–151. MIT Press, Cambridge, MA, États-Unis, 1998.

[HG96] D. Hull and G. Grefenstette. A Detailed Analysis of English Stemming Algorithms. Technical Report, Xerox Research Centre Europe, Meylan, France, 1996.

[HGS+97] D. Hull, G. Grefenstette, B.M. Schulze, H. Schütze, and J.O. Pedersen. Xerox TREC-5 Site Report: Routing, Filtering, NLP and Spanish Tracks. In *Proceedings of the 5th Text Retrieval Conference, TREC*, Gaithersburg, États-Unis, 1997.

[Hul96] D. Hull. Stemming Algorithms - A Case Study for Detailed Evaluation. *Journal of the American Society of Information Science*, 47(1):70–84, 1996.

[Jac94] C. Jacquemin. FASTR: A Unification-Based Front-End to Automatic Indexing. In *Proceedings of Intelligent Multimedia Information Retrieval Systems and Management, 4th International Conference Recherche d'Informations Assistée par Ordinateur, RIAO 94*, New York, États-Unis, 1994.

[JC94] Y. Jing and W.B. Croft. An Association Thesaurus for Information Retrieval. In *Proceedings of the 4th International Conference Recherche d'Informations Assistée par Ordinateur, RIAO 94*, New York, États-Unis, 1994.

[JKT97] C. Jacquemin, J.L. Klavans, and E. Tzoukermann. Expansion of Multi-Word Terms for Indexing and Retrieval Using Morphology and Syntax. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, ACL*, Madrid, Espagne, 1997.

[Jou95] C. Jouis. Seek, un logiciel d'acquisition de connaissances utilisant un savoir linguistique sans employer de connaissances sur le monde externe. In *Proceedings of 6es Journées Acquisition, Validation, JAVA'95*, Grenoble, France, 1995.

[Kho95] C. S.G. Khoo. *Automatic Identification of Causal Relations in Text and their Use for Improving Precision in Information Retrieval.* PhD thesis, Université de Syracuse, New-York, États-Unis, 1995.

[KP98] W. Kraaij and R. Pohlmann. Comparing the Effect of Syntactic vs. Statistical Phrase Indexing Strategies for Dutch. In C. Nicolaou and C. Stephanides, editors, *Proceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries, ECDL'98*, volume 1513, pages 605–614. Lecture Notes in Computer Science, Springer Verlag, Berlin, Heidelberg, Allemagne, 1998.

[Kro93] R. Krovetz. Viewing Morphology as an Inference Process. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Pittsburgh, États-Unis, 1993.

[Lar98] Larousse. *Le Petit Larousse 1999.* Larousse, Paris, France, 1998.

[LBEM98] C. Loupy, P. Bellot, M. El Bèze, and P.-F. Marteau. Query Expansion and Classification of Retrieved Documents. In *Proceedings of the 7th Text Retrieval Conference, TREC*, Gaithersburg, États-Unis, 1998.

[LC90] D.D. Lewis and W.B. Croft. Term Clustering of Syntactic Phrases. In *Proceedings of the 13th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Bruxelles, Belgique, 1990.

[Lew92] D.D. Lewis. An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Copenhague, Danemark, 1992.

[Lov68] J.B. Lovins. Development of a Stemming Algorithm. *Mechanical Translation and Computational Linguistics*, 11:22–31, 1968.

[LPTW81] M. Lennon, D.S. Pierce, B.D. Tarry, and P. Willett. An Evaluation of some Conflation Algorithms for Information Retrieval. *Journal of Information Science*, 3(1):177–183, 1981.

[MBSC97] M. Mitra, C. Buckley, A. Singhal, and C. Cardie. An Analysis of Statistical and Syntactic Phrases. In *Proceedings of the 5th International Conference Recherche d'Informations Assistée par Ordinateur, RIAO 97*, Montréal, Canada, 1997.

[MM00a] R. Mihalcea and D.I Moldovan. Semantic Indexing Using WORDNET Senses. In *Proceedings of ACL Workshop on IR & NLP*, Hong Kong, 2000.

[MM00b] D.I. Moldovan and R. Mihalcea. Using WORDNET and Lexical Operators to Improve Internet Searches. *IEEE Internet Computing*, 4(1):34–43, 2000.

[MML00] I. Moulinier, J.A. McCulloh, and E. Lund. West Group at CLEF 2000: Non-English Monolingual Retrieval. In *Proceedings of the Workshop of Cross-Language Evaluation Forum, CLEF 2000*, Lisbonne, Portugal, 2000.

[MTT99] R. Mandala, T. Tokunaga, and H. Tanaka. Combining Multiple Evidence from Different Types of Thesaurus for Query Expansion. In *Proceedings of the 22th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, États-Unis, 1999.

[Nam00] F. Namer. FLEMM : un analyseur flexionnel du Français à base de règles. *Traitement Automatique des Langues*, 41(2):523–547, 2000.

[NL02] H.T. Ng and Y.T. Lee. An Empirical Evaluation of Knowledge Sources and Learning Algorithms for Word Sense Disambiguation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 02*, Philadelphie, Penn., États-Unis, 2002.

[Pai94] C.D. Paice. An Evaluation Method for Stemming Algorithms. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Dublin, Irlande, 1994.

[PB97] T. Pedersen and R. Bruce. A New Supervised Learning Algorithm for Word Sense Disambiguation. In *Proceedings of the 14th National Conference on Artificial Intelligence, AAAI*, Providence, États-Unis, 1997.

[Pol03] A. Polguère. *Lexicologie et sémantique lexicale - Notions fondamentales.* Presse de l'Université de Montréal, Montréal, Québec, Canada, 2003.

[Por80] M.F. Porter. An Algorithm for Suffix Stripping. *Program*, 14:130–137, 1980.

[PS00] R. Pichon and P. Sébillot. From Corpus to Lexicon: From Contexts to Semantic Features. In B. Lewandowska-Tomaszczyk and P.J. Melia, editors, *PALC'99: Practical Applications in Language Corpora*, volume 1 of *Lodz studies in Language*, pages 375–389. Peter Lang, 2000.

[Pus95] J. Pustejovsky. *The Generative Lexicon.* Cambridge: MIT Press, 1995.

[PW91] H.J. Peat and P. Willett. The Limitations of Term Cooccurrence Data for Query Expansion in Document Retrieval Systems. *Journal of the American Society for Information Science*, 42(5):378–383, 1991.

[PW92] M. Popovic and P. Willett. The Effectiveness of Stemming for Natural-Language Access to Slovene Textual Data. *Journal of the American Society for Information Science*, 43(5):384–390, 1992.

[QF95] Y. Qiu and H.P. Frei. Improving the Retrieval Effectiveness by a Similarity Thesaurus. Rapport interne 225, Department of Computer Science, ETH Zürich, Zürich, Suisse, 1995.

[RBC00] M. Rajman, R. Besançon, and J.-C. Chappelier. Le modèle DSIR : une approche à base de sémantique distributionnelle pour la recherche documentaire. *Traitement automatique des langues*, 41(2):549–578, 2000.

[Roc71] J.J. Rocchio. Relevance Feedback in Information Retrieval. In G. Salton, editor, *The SMART Retrieval System: Experiments in Automatic Document Processing*, pages 313–323. Prentice-Hall, Englewood Cliffs, NJ, États-Unis, 1971.

[RS03] M. Rossignol and P. Sébillot. Extraction statistique sur corpus de classes de mots-clés thématiques. *Traitement automatique des langues*, 44(3):217–246, 2003.

[Sag81] N. Sager. *Natural Language Information Processing: A Computer Grammar of English and its Applications.* Addison-Wesley, Reading, Mass., États-Unis, 1981.

[San97] M. Sanderson. *Word Sense Disambiguation and Information Retrieval.* PhD thesis, Université de Glasgow, Écosse, 1997.

[Sav93] J. Savoy. Stemming of French Words Based on Grammatical Categories. *Journal of the American Society for Information Science*, 44(1):1–9, 1993.

[Sch72] R.C. Schank. Dependency: a Theory of Natural Language Understanding. *Cognitive Psychology*, 3(4):532–631, 1972.

[SLWPC99] T. Strzalkowski, F. Lin, J. Wang, and J. Perez-Carballo. Evaluating Natural Language Processing Techniques in Information Retrieval. In T. Strzalkowski, editor, *Natural Language Information Retrieval*, pages 113–145. Kluwer Academic Publishers, 1999.

[Sme99] A.F. Smeaton. Using NLP or NLP Resources for Information Retrieval Tasks. In T. Strzalkowski, editor, *Natural Language Information Retrieval*, pages 99–111. Kluwer Academic Publishers, 1999.

[Sow84] J.F. Sowa. *Conceptual Structures: Information Processing in Mind and Machine.* Addison-Wesley, Reading, Mass., États-Unis, 1984.

[SP95] H. Schütze and J.O. Pedersen. Information Retrieval Based on Word Senses. In *Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, États-Unis, 1995.

[SP97] H. Schütze and J.O. Pedersen. A Cooccurrence-Based Thesaurus and two Applications to Information Retrieval. *Information Processing & Management*, 33(3):307–318, 1997.

[Spa99] K. Sparck Jones. What is the Role of NLP in Text Retrieval? In T. Strzalkowski, editor, *Natural Language Information Retrieval*, pages 1–24. Kluwer Academic Publishers, 1999.

[SQ96] A. Smeaton and I. Quigley. Experiments on Using Semantic Distances Between Words in Image Caption Retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Zürich, Suisse, 1996.

[ST84] K. Sparck Jones and J.I. Tait. Automatic Search Term Variant Generation. *Journal of Documentation*, 40(1):50–66, 1984.

[SYY75] G. Salton, C.S. Yang, and C.T. Yu. A Theory of Term Importance in Automatic Text Analysis. *Journal of the American Society for Information Science*, 26(1):33–44, 1975.

[THH00] J. Takenobu, O. Hironori, and T. Hozumi. Effectiveness of Complex Index Terms in Information Retrieval. In *Proceedings of the 6th International Conference Recherche d'Informations Assistée par Ordinateur, RIAO 2000*, Paris, France, 2000.

[TV98] G.G. Towell and E.M. Voorhees. Disambiguating Highly Ambiguous Words. *Computational Linguistics*, 24(1):125–145, 1998.

[UKY99] O. Uzuner, B. Katz, and D. Yuret. Word Sense Disambiguation for Information Retrieval. In *Proceedings of the 16th National Conference on Artificial Intelligence, AAAI*, Orlando, États-Unis, 1999.

[Van91] C.J. Van Rijsbergen. Towards an Information Logic. In *Proceedings of the 12th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Cambridge, États-Unis, 1991.

[VBA02] J. Vilares Ferro, F.M. Barcala, and M.A. Alonso. Using Syntactic Dependency-Pairs Conflation to Improve Retrieval Performance in Spanish. In *Proceedings of the 3th International Conference on Intelligent Text Processing and Computational Linguistics,CICLING*, Mexico, Mexique, 2002.

[Voo98] E.M. Voorhees. Using WORDNET for Text Retrieval. In C. Fellbaum, editor, WORDNET*: An Electronic Lexical Database*. The MIT Press, 1998.

[WA98] W.A. Woods and J. Ambroziak. Natural Language Technology in Precision Content Retrieval. In *Proceedings of the International Conference on Natural Language Processing and Industrial Applications, NLP+IA 98*, Moncton, Canada, 1998.

[Wal93] P. Wallis. Information Retrieval Based on Paraphrase. In *Proceedings of the 1st Pacific Association for Computational Linguistics Conference, PACLING*, Vancouver, Canada, 1993.

[WFG+90] Y. Wilks, D. Fass, C.M. Guo, J.E. McDonald, T. Plate, and B.M. Slator. Providing Machine Tractable Dictionary Tools. *Machine Translation*, 5(2):99–154, 1990.

[XC98] J. Xu and W.B. Croft. Corpus-Based Stemming Using Cooccurrence of Word Variants. *ACM Transactions on Information Systems*, 16(1):61–81, 1998.

[XC00] J. Xu and W.B. Croft. Improving the Effectiveness of Information Retrieval with Local Context Analysis. *ACM Transactions on Information Systems*, 18(1):79–112, 2000.

[Yar95] D. Yarowsky. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *Proceedings of the 33th Annual Meeting of the Association for Computational Linguistics, ACL 95*, Cambridge, Mass., États-Unis, 1995.

[Yar00] D. Yarowsky. Hierarchical Decision List for Word Sense Disambiguation. *Computers and the Humanities*, 34(1–2):179–186, 2000.

[ZGD01] P. Zweigenbaum, N. Grabar, and S. Darmoni. Apport de connaissances morphologiques pour la projection de requêtes sur une terminologie normalisée. In *Proceedings of 8ème conférence annuelle sur le Traitement automatique des langues naturelles, TALN'01*, Tours, France, 2001.

[ZM94] P. Zweigenbaum and Consortium Menelas. MENELAS: An Access System for Medical Records Using Natural Language. *Computer Methods and Programs in Biomedicine*, 45:117–120, 1994.

[ZTME97] C. Zhai, X. Tong, N. Milic-Frayling, and D.A. Evans. Evaluation of Syntactic Phrase Indexing - CLARIT NLP Track Report. In *Proceedings of the 5th Text Retrieval Conference, TREC*, Gaithersburg, États-Unis, 1997.

Unité de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unité de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unité de recherche INRIA Rhône-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN
Unité de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unité de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
http://www.inria.fr
ISSN 0249-6399